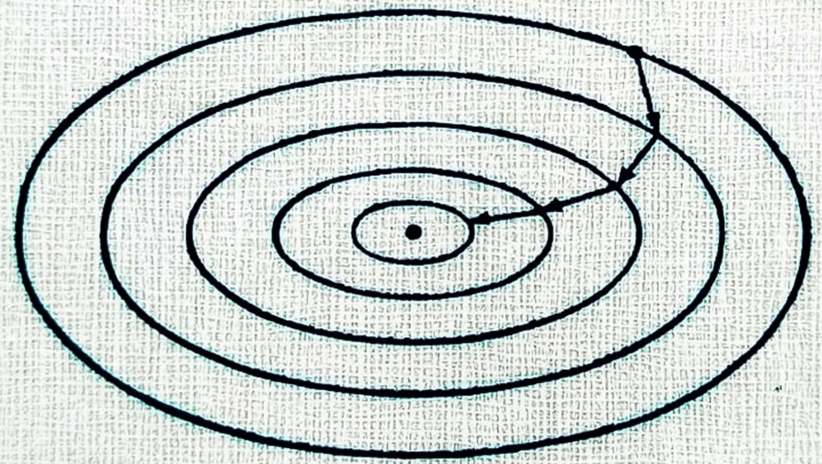
V. G. Karmanov

MATHEMATICAL PROGRAMMING



MIR PUBLISHERS
MOSCOW

V.G. Karmanov MATHEMATICAL PROGRAMMING

MATHEMATICAL PROGRAMMING

В. Г. Карманов

МАТЕМАТИЧЕСКОЕ ПРОГРАММИРОВАНИЕ

Издательство «Наука» Москва

V.G.Karmanov

MATHEMATICAL PROGRAMMING

Mir Publishers
Moscow

Translated from the Russian
by Vladimir Volosov

First published 1989
Revised from the 1986 Russian edition

TO THE READER

Mir Publishers would be grateful for your comments on the content, translation and design of this book. We would also be pleased to receive any other suggestions you may wish to make.

Our address is:
Mir Publishers
2 Pervy Rizhsky Pereulok
I-110, GSP, Moscow, 129820
USSR

На английском языке

Printed in the Union of Soviet Socialist Republics

ISBN 5-03-000533-1

© Издательство «Наука».
Главная редакция физико-математической литературы, 1975; с изменениями, 1980; с добавлениями, 1986

© English translation, Mir Publishers, 1989

CONTENTS

Preface 8

Chapter 1. Introduction 10

1.1. Subject Matter of Mathematical Programming 10
1.2. On Models 13
1.3. Classification and Specificity 13
1.4. Examples of Mathematical Models 15
1.5. Basic Notation 21

Chapter 2. Elements of Convex Analysis 24

2.1. Euclidian Space. Convex Sets 24
2.2. Projection. Separation Theorems 26
2.3. Cone. Farkas Theorem 33
2.4. Convex Functions 37
2.5. Strongly Convex Functions 44

Chapter 3. Fundamentals of Mathematical Programming 48

3.1. Mathematical Programming Problems 48
3.2. Feasible Directions 49
3.3. Extremal Properties 52
3.4. Extremal Properties on Convex Sets 54
3.5. Sufficient Conditions for Optimality 58
3.6. Lagrangian Function. Optimality Conditions 60

Chapter 4. Linear Programming Theory 68

4.1. Basic Notions 68
4.2. Basic Theorems 70
4.3. Algebraic Characterization of an Extreme Point 79
4.4. Dual Problems with Mixed Constraints 82
4.5. Canonical Form of a Linear Programming Problem 85

Chapter 5. Finite Methods of Solving Linear Programming Problems 88

5.1. Simplex Method 88
5.2. Recurrence Relations of the Simplex Method Algorithm (Relationship between the Parameters of Successive Iterations) 93

5.3. Methods of Finding an Initial Extreme Point 94
5.4. Degeneracy. Perturbation Method 98
5.5. On the Simplex Method for Solving Special Classes of Linear Programming Problems 101
5.6. On the Modified Simplex Method 102
5.7. Multiplicative Representation of an Inverse Matrix 103
5.8. Dual Simplex Method 104
5.9. Solution of the Dual Problem Interpreted as Estimation of Influence 107
5.10. On the Dual Simplex Method for Problems with Increasing Number of Constraints 109
5.11. Method of Simultaneous Solution of the Primal and Dual Problems 111
5.12. Decomposition Method 116

Chapter 6. Penalty Function Method 122

6.1. Description of the Method 122
6.2. Convergence Theorems 126

Chapter 7. Stability Problems in Mathematical Programming 137

7.1. Correctly and Incorrectly Posed Problems 137
7.2. A Class of Correct Problems 140
7.3. Problems with Exact Constraints. Regularization Method 142
7.4. Convergence 143
7.5. Regularization Method (the General Case) 145
7.6. Convergence of the Regularization Method (the General Case) 146

Chapter 8. One-Dimensional Minimization Methods 151

8.1. On One-Dimensional Minimization Problems 151
8.2. Search for an Interval Containing the Point of Minimum 154
8.3. Fibonacci and Golden Section Methods 154
8.4. Parabolic Method 163
8.5. Cubic Approximation Method 164
8.6. Method of Tangents 165

Chapter 9. Relaxation Methods of Solving Extremal Problems. Unconditional Minimization Methods 171

9.1. Convergence and Stability Problems. Relaxation Processes 171
9.2. Auxiliary Material 173
9.3. Theorems on Estimates 182
9.4. Descent Methods 186
9.5. First- and Second-Order Methods 198
9.6. Method of Conjugate Directions 209
9.7. Zero-Order Methods 214

Chapter 10. Relaxation Methods of Solving Extremal Problems with Constraints 222

10.1. Description of the Methods 222
10.2. Gradient Projection Method 224
10.3. Method of Conditional Gradient 229
10.4. Method of Feasible Directions 234
10.5. Methods of Finding a Feasible Point 244
10.6. Methods of Random Descent 249

Chapter 11. Modified Lagrange Function Method 269

11.1. Lagrange Multiplier Method 269
11.2. Modified Lagrange Function Method 270
11.3. Interrelation between the Lagrange Multiplier Method and the Penalty Function Method 276

Supplements 280
Bibliography 301
Name Index 302
Subject Index 303

PREFACE

This book presents basic concepts of mathematical programming theory and numerical methods of solving the corresponding extremal problems.

It is based on lectures delivered by the author for a number of years at the Department of Mechanics and Mathematics and the Department of Computational Mathematics and Cybernetics of Moscow University.

Naturally, in a course meant for an initial acquaintance with the subject only basic numerical methods are given. Therefore this is not a reference book where the reader can find a specific algorithm required for solving an actual optimization problem.

Chapter 1 is devoted to the subject of mathematical programming, its role in the operations research, and examples of technical and economic applications whose mathematical models are problems of mathematical programming.

Chapter 2 presents mathematical background for the subsequent chapters of the book.

In Chapter 3 emphasis is laid upon optimality conditions—necessary conditions for the existence of local minimum in nonlinear programming problems, and necessary and sufficient conditions for solvability of convex programming problems.

Chapter 4 deals with linear programming theory.

Basic methods of solving linear programming problems are presented in Chapter 5.

The widespread penalty function method is considered in Chapter 6.

In Chapter 7 the important notion of stability of linear programming problems is introduced. Here primary attention is placed on the regularization method for solving unstable extremal problems.

Repeated solution of one-dimensional minimization problems is a necessary stage in the majority of numerical

linear programming methods, and this brought into being Chapter 8.

At first glance, methods of solving unconditional optimization problems belong to "numerical analysis"; however, these methods have become an essential part of mathematical programming. For instance, extremal problems with constraints (conditional extremum problems) are reduced by means of the penalty function method to unconditional optimization problems, and the same relates to the regularization and modified Lagrange function methods. Moreover, the ideas of many of the methods of solving conditional extremum problems are based on unconditional optimization methods given in Chapter 9.

Chapter 10 is devoted to methods of solving extremal problems with constraints: the gradient projection, conditional gradient, feasible direction, and random descent methods.

In recent years specialists have focused their attention on methods of solving mathematical programming problems based on the construction of the modified Lagrange functions. These questions are treated in Chapter 11.

In the investigation of the methods presented in this book primary emphasis is placed on convergence and stability problems and also on a priori estimates for the rate of convergence, one of the important characteristics of the efficiency of the methods.

In relation to numerical methods the following remark should be made: when solving a real problem, even a method which at first glance seems the most disadvantageous may then turn out to be extremely effective; in this connection the author refrains from the presentation of comparative characteristics of the methods lying beyond the estimates for the rate of convergence.

The book is concluded by supplements presenting some results which may prove useful to readers who wish to study mathematical programming more thoroughly.

The author expresses his deep gratitude to V.A. Bereznev, D.V. Denisov, and A.A. Tret'yakov for valuable advice, critical comment, and help.

V.G. Karmanov

Chapter 1

INTRODUCTION

1.1. Subject Matter of Mathematical Programming

1.1.1. The content of mathematical programming. Mathematical programming consists of the theory and methods for solving problems of finding extrema of functions on sets determined by linear and nonlinear constraints (in the form of equations and inequalities). Mathematical programming is one of the divisions of operations research.

1.1.2. Applications. Mathematical programming problems have extensive applications in various fields of human activity where it is necessary to choose one of the possible ways (programs) of action, for instance, in controlling and planning industrial processes, designing, long-term planning, military engineering, etc.

A considerable number of problems arising in human society is related to controllable phenomena, i.e. phenomena that can be governed by working out elaborated decisions. With the limited amount of information available at early stages of the development of human society, an optimal, in a sense, decision was made on the basis of intuition and experience, and then, as the amount of information about the phenomenon under study increased, by using direct calculations. For instance, this was the case in production planning.

A completely different situation appears in a modern industry with multiserial and multirange production where the amount of input information is so extensive that its processing in order to make a definite decision is impossible without using modern electronic computers. Still greater difficulties arise in relation to the problem of making the best decision.

1.1.3. Decision making. In operations research by decision making is meant a complex process in which the following four basic stages can be indicated.

The 1st stage. Constructing a qualitative model of the problem in question, i.e. separating out the factors that seem the most important and establishing regularities they obey.

This stage usually lies outside the framework of mathematics.

The 2nd stage. Constructing a mathematical model for the problem, i.e. expressing the qualitative model in mathematical terms. Thus, a mathematical model is an abstruct representation of a real phenomenon written in mathematical symbols and constructed in such a way that its analysis provides a possibility of understanding the essence of the phenomenon. A mathematical model establishes relations between the variables—the control parameters of the phenomenon.

This stage also includes the construction of an *objective function*, i.e. a numerical characteristic to whose maximum (or minimum) value there corresponds a situation which is the best from the viewpoint of the decision maker.

The second stage requires a knowledge of mathematics.

The result of the two stages is a statement of the corresponding mathematical problem.

The 3rd stage. Investigating the dependence of the value of the objective function on the variables.

This stage requires application of the mathematical apparatus meant for solving mathematical problems arising at the second stage of the decision making process.

A wide class of control problems are extremal problems whose mathematical models involve constraints on the variables expressed by equations and inequalities. It is the theory and methods of solving these problems that constitute the content of mathematical programming.

At the third stage the solution to the corresponding extremal problems is found using the mathematical apparatus. It should be noted that, as a rule, mathematical programming problems applied to practical problems contain a very large number of variables and constraints. The amount of calculations required for finding the corresponding solutions is so large that the whole process can be realized only by modern computers and thus requires either the elaboration of special computer

programs for certain algorithms or the application of the available standard programs.

The 4th stage. The comparison of the calculation results obtained at the 3rd stage with the modelled object, i.e. an expert examination of the results (the practice criterion).

Hence, at this stage the degree of adequacy of the model to the modelled object is estimated to within the accuracy of the initial information.

Here the following two cases are possible.

The 1st case. If the results of the comparison are unsatisfactory (which is a usual situation at the initial stage of a modelling process), one passes to the second cycle of the process: the input information about the modelled object is specified and, when necessary, the setting of the problem (the 1st stage) is reconsidered, the mathematical model is refined or renewed (the 2nd stage), the corresponding mathematical problem is solved (the 3rd stage), and, finally, the comparison is carried out again (the 4th stage).

The 2nd case. If the results of the comparison are satisfactory, the model is accepted. When the calculation results are meant to be used many times, it becomes necessary to prepare the model for operation. For example, assume that the purpose of the modelling is the production planning. Then the operation of the model includes information accumulation and processing, introduction of the processing results into a computer, calculation of production planning based on the elaborated programs, and, finally, representation of the calculation results (in a form convenient for the user) for their application in the industrial process.

The preparation of the model for operation implies working out a special mathematical software without which the model cannot be used for practical purposes: it is necessary to create a flexible system of programs allowing a user with ordinary mathematical qualification to get in contact with a computer.

At the same time, the mathematical software as well as the model itself must admit of modernization in view of the new requirements that are constantly imposed on the industry. We stress that it is modernization and not

the creation of new systems of programs that is meant here. Only under these conditions can mathematical models and computers be regularly used in control processes.

1.2. On Models

Since a course in mathematical programming includes the proofs of a great number of various theorems and the consideration of versatile methods of solving extremal problems, one may think that the role of a mathematician in solving applied problems is confined to the third stage of the modelling process whereas in other stages only specialists familiar with non-formalized aspects of the modelled object are involved. This is sometimes the case, but as a rule the results of such division of labour in the modelling process prove unsatisfactory. In fact, the model represents the properties of the object only approximately, and the accuracy of the approximation must be compatible with that of the input information about the phenomenon.

One often tries to construct a mathematical model comprising a great amount of input data. However, a thorough analysis shows that due to their insufficient accuracy a considerable part of the data either slightly influence the solution or have no effect on it at all. Moreover, it should be taken into consideration that a mathematical model with a great number of parameters leads to an optimization problem for functions with so large a number of variables that an attempt to find its numerical solution with an accuracy compatible with the input information proves hopeless.

Thus, there is a number of conditions imposed on mathematical models that make it necessary for mathematicians to take part in the solution of applied problems as early as the second stage of the modelling process.

1.3. Classification and Specificity

1.3.1. Basic divisions. In mathematical programming it is possible to separate out two basic divisions. The first and quite thoroughly developed is the mathematical

programming proper. It comprises deterministic problems where the whole input information is completely defined.

The other division, the so-called stochastic programming, deals with problems where the input information involves elements of indeterminacy or some of the parameters of the problem are random variables with known probability characteristics. For instance, in industrial planning the information about the real situation in which the elaborated plan is to be realized is often incomplete. Another example is an extremal problem, say, for modelling an automatic device whose operation is accompanied by random noise. It should be noted that one of the main difficulties encountered in stochastic programming lies in the very setting of the problems primarily owing to the complexity of the analysis of the initial information.

1.3.2. Classification of mathematical programming problems. Traditionally, mathematical programming problems are classified in the following basic types:

Linear programming problems where the objective function is linear and the set on which an extremum of the objective function is sought is determined by a system of linear equations and inequalities. Further, in linear programming there are classes of problems whose structure admits of working out special methods for their solution that have some advantages over the methods of solving general problems. For instance, such is the class of *transportation problems.*

Nonlinear programming problems where both the objective function and the constraints are nonlinear. Nonlinear programming problems are usually subdivided in the following way:

Convex programming problems where the objective function is convex (when the problem of its minimization is considered) and the set on which the extremal problem is solved is also convex.

Quadratic programming problems where the objective function is quadratic and the constraints are linear equations and inequalities.

Multiextremum problems. Here specialized classes of problems often encountered in applications are usually

considered, e.g. problems of minimizing concave functions on a convex set.

An important class is formed by *integer programming problems* where the condition that the variables assume integral values is imposed.

1.3.3. Specificity. What is the specificity of mathematical programming problems?

As a rule, mathematical programming problems do not admit of application of classical methods of finding conditional extrema since even in the simplest case of linear programming problems the extrema are attained at extreme points of the feasible region, i.e. points where the differentiability is violated. The most powerful method of solving extremal problem in classical differential calculus—the Lagrange multiplier method—is elaborated for the case when the set of conditions of the problem is specified by a system of equations and not by a system of inequalities.

Another specific property is that in practical problems the number of variables and constraints is so large that if the points "suspected of extremality" (for example, all the extreme points) are simply tested consecutively, no modern computer can solve this problem in a reasonable time.

In view of what has been said, the aim of mathematical programming is to create, when possible, analytical methods for determining solutions and, when there are no such methods, to elaborate effective computation means for finding an approximate solution.

Finally, we note that the name "mathematical programming" stems from the fact that the aim of the solution of the problem is to work out a program of action.

1.4. Examples of Mathematical Models

The theory of mathematical models is the subject matter of a special course whose primary requirement is that the reader be familiar with the branch of knowledge the modelled object belongs to.

In this book we present traditional examples demonstrating the application of the mathematical modelling method in economic problems.

1.4.1. The diet problem. Given different foods with a known content of nutrients in each of them and the cost of the foods, it is required to work out a full value diet with minimum cost and satisfying the necessary nutritional requirements.

Let there be n different foods and m nutrients (e.g. fat, proteins, carbohydrates, vitamins, etc.). Denote by a_{ij} the content (in units of mass) of the jth nutrient in the unit mass of the ith food and by b_j the minimum (in units of mass) daily need for the jth nutrient. Finally, let x_i denote the sought-for daily consumption of the ith food. Obviously, $x_i \geqslant 0$.

The sum $\sum_{i=1}^{n} a_{ij}x_i$ is the total content of the jth nutrient in the diet, and it must not be less than the minimum need b_j:

$$\sum_{i=1}^{n} a_{ij}x_i \geqslant b_j, \quad j = \overline{1, m}.$$

If c_i is the cost of the unit mass of the ith food, then the cost of the whole diet is given by the linear form $\sum_{i=1}^{n} c_i x_i$.

Thus, the mathematical statement of the diet problem is: find

$$\min \sum_{i=1}^{n} c_i x_i$$

under the conditions

$$\sum_{i=1}^{n} a_{ij}x_i \geqslant b_j, \quad j = \overline{1, m},$$

$$x_i \geqslant 0, \quad i = \overline{1, n}.$$

This is one of the typical linear programming problems.

It is quite clear that the same statement as for the diet problem applies to various problems of optimal mixtures (here the term "mixture" should be understood in the generalized sense including a mixture proper, an alloy, a diet, etc.).

1.4.2. Transportation problem. Another typical example of a linear programming problem is the *transportation*

problem. It is required to work out a transportation plan for the delivery of a certain product so that the total cost of the transportation be minimum.

The initial information:

a_i is the amount of the product (measured in certain units) at the ith source ($i = \overline{1, m}$);

b_j is the demand for the product (expressed in the same units) at the jth destination ($j = \overline{1, n}$);

c_{ij} is the cost of the transportation of the unit of the product from the ith source to the jth destination.

Denote by x_{ij} the amount of the product to be transported from the ith source to the jth destination.

According to this notation:

$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}x_{ij}$ is the total cost of the transportation;

$\sum_{j=1}^{n} x_{ij}$ is the amount of the product transported from the ith source;

$\sum_{i=1}^{m} x_{ij}$ is the amount of the product transported to the jth destination.

In the simplest case the following obvious conditions must hold:

$$\sum_{i=1}^{n} x_{ij} = a_i, \quad i = \overline{1, m},$$

$$\sum_{j=1}^{m} x_{ij} = b_j, \quad j = \overline{1, n},$$

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j.$$

The mathematical formulation of the transportation problem reads: find

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij}x_{ij}$$

under the conditions

$$\sum_{j=1}^{n} x_{ij} = a_i,$$

$$\sum_{i=1}^{m} x_{ij} = b_j,$$

$$x_{ij} \geqslant 0, \quad i = \overline{1, m}, \quad j = \overline{1, n}.$$

This is called the *closed* transportation problem.

We note that the relation

$$\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$$

is a natural condition for solvability of the closed transportation problem.

A more general statement specifies the so-called *open* transportation problem: find

$$\min \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij}$$

under the conditions

$$\sum_{j=1}^{n} x_{ij} \leqslant a_i,$$

$$\sum_{i=1}^{m} x_{ij} = b_j,$$

$$x_{ij} \geqslant 0, \quad i = \overline{1, m}, \quad j = \overline{1, n}.$$

It is clear that in this problem it is not implied that the whole amount of product piled up in the ith source must be taken out.

The scheme of the transportation problem also embraces some other technical and economic problems, e.g. the so-called assignment problem on the most economical (in the sense of optimal time) distribution of n kinds of jobs between n workers, given the aptitude of worker i to job j. This problem is a special case of the closed transportation problem for $m = n$ and $a_i = b_j = 1$.

We note that the solutions to transportation problems are integers and therefore these problems belong to the class of integer linear programming problems where the variables have to be integer-valued (an additional necessary condition).

1.4.3. The problem on the operation regime of a power system. As an example of a convex programming problem we consider the simplest version of a problem on the optimal operation of a power system.

Let there be an isolated power system consisting of thermal power stations connected by transmission lines to a terminal where the load is concentrated. We examine the distribution of the active power between the power stations at a given time. The distribution is performed according to the criterion of minimum total fuel consumption in the generation of the active power.

Denote by x_i the active power generated at the ith station. The values x_i lie between α_i and β_i determined by technical conditions: $\alpha_i \leqslant x_i \leqslant \beta_i$. Besides, the condition of power balance must be fulfilled, i.e. the total generated power must equal the consumed power P plus the total loss of power π in the transmission lines:

$$\sum_{i=1}^{m} x_i = P + \pi.$$

The fuel consumption in the generation of the power x_i is a convex function $T_i(x_i)$ on the closed interval $[\alpha_i, \beta_i]$.

Thus, the problem is stated in the following way: find

$$\min \sum_{i=1}^{m} T_i(x_i)$$

under the conditions

$$\sum_{i=1}^{m} x_i = P + \pi,$$

$$\alpha_i \leqslant x_i \leqslant \beta_i, \quad i = \overline{1, m}.$$

The constructed model is a typical *convex programming problem with linear constraints*. The solution to the problem provides a rather crude approximation to a real optimal operation regime of a power system. In a real situation it is incorrect to assume that the whole load is concentrated at a single terminal and it is necessary to consider n terminals. Moreover, the losses in the system are not constant and depend on the magnitudes of the transmitted powers and the transmission line parameters.

As the next approximation one can consider a problem where π is a bilinear function in x_{ij} ($i = \overline{1, m}$, $j = \overline{1, n}$), the control parameter x_{ij} being the magnitude of the active power transmitted from the ith power station to the jth terminal.

It is obvious that in the new model the conditions involve nonlinear functions (the losses $\pi(x_{ij})$ in the balance equation).

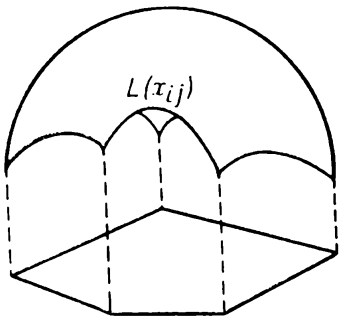


Fig. 1.1

This is also a *convex programming problem*, but of a more complex type than the previous one.

An example of a multiextremum problem is the simplest *allocation problem.*

1.4.4. Allocation problem. There are m places of consumption of a product ($1, 2, \ldots, j, \ldots, m$) with a given volume of consumption b_j at each place. There are also n possible places of production ($1, 2, \ldots, i, \ldots, n$), the dependence of the production cost f_i on the output x_i being known for each ith place. (It is assumed that the production cost $f_i(x_i)$ includes capital outlays.) Finally, the transportation cost matrix a_{ij} is known (a_{ij} is the cost of the transport of the unit product from the ith production place to the jth consumption place). It is required to determine the volumes of transportation x_{ij} from the ith production place to the jth consumption place and the outputs $x_i = \sum_j x_{ij}$ so that the total cost is minimized; in other words, it is required to find

$$\min \left[L(x_{ij}) = \sum_{i,\, j} a_{ij} x_{ij} + \sum_i f_i(x_i) \right]$$

under the conditions

$$\sum_i x_{ij} = b_j, \quad x_{ij} \geqslant 0.$$

Since, as a rule, the cost of the unit product usually decreases with the output growth, the functions $f_i(x_i)$ increase monotonically and are convex upward. The set of the values of x_{ij} satisfying the constraints of the

problem forms a convex polyhedron whose vertices are points of local minima of the function $L(x_{ij})$ (Fig. 1.1). This accounts for the name *multiextremum* problems.

1.5. Basic Notation

E_n is the n-dimensional Euclidean space.

$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$ is a column vector.

$\mathbf{x}^T = (x_1, x_2, \ldots, x_n)$ is a row vector.

$\mathbf{x} \in R$ means that $\mathbf{x}$ belongs to R.

$\mathbf{x} \notin R$ means that $\mathbf{x}$ does not belong to R.

$R \subset S$ means that R is contained in S (a set-theoretical relation).

$R \cap S$ is the set-theoretical intersection of the sets R and S.

$\bigcap_{\lambda \in \Lambda} R_\lambda$ is the set-theoretical intersection of the sets R_λ over the index λ ($\lambda \in \Lambda$).

$R \cup S$ is the set-theoretical union of the sets R and S.

$\bigcup_{\lambda \in \Lambda} R_\lambda$ is the set-theoretical union of the sets R_λ over the index λ.

$\varnothing$ is an empty set.

$G(R)$ is the boundary of the set R.

$\{\mathbf{x}: Q\}$ is the set of all elements $\mathbf{x} \in E_n$ possessing the property Q.

$\{\mathbf{x} \in Y: Q\}$ is the set of all elements $\mathbf{x} \in Y$ possessing the property Q.

$R \setminus S$ is the set-theoretical difference, i.e. the set $\{x: x \in R, x \notin S\}$.

$\operatorname{int} R = R \setminus G(R)$ is the interior of the set R.

$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i$ is the scalar product of the vectors $\mathbf{x}$ and $\mathbf{y}$.

$\| \mathbf{x} \| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$ is the Euclidean norm of the vector $\mathbf{x}$.

$\{\mathbf{x}_k\}$ is a sequence of points $\mathbf{x}_k$.

$\{\alpha_k\}$ is a sequence of numbers α_k.

$A = [a_{ij}]$ is an $m \times n$ matrix.

$i = \overline{1, n}$ means that the index i runs over all integers from 1 to n.

A^T is the transpose of a matrix A: $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ where

$$\mathbf{a}_i = \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{mi} \end{bmatrix}, \quad i = \overline{1, n}, \quad A^T = \begin{bmatrix} \mathbf{a}_1^T \\ \mathbf{a}_2^T \\ \vdots \\ \mathbf{a}_n^T \end{bmatrix}.$$

AB is the product of the matrices A and B:

$$C = [c_{ij}] = AB = \left[\sum_{k=1}^{n} a_{ik} b_{kj} \right].$$

B^{-1} is the inverse matrix of a square matrix B.

$(A\mathbf{x})_i$ is the ith component of the vector $A\mathbf{x}$.

$\det B$ is the determinant of a square matrix B.

$\mathbf{x} \geqslant \mathbf{y}$ is the partial ordering in E_n meaning that $x_i \geqslant y_i$ $(i = \overline{1, n})$.

$\mathbf{x} > \mathbf{y}$ means that $x_i > y_i$ $(i = \overline{1, n})$.

$\mathbf{x} \neq \mathbf{y}$ means that $x_i \neq y_i$ for at least one index i.

$[\mathbf{x}, \mathbf{y}]$ is a closed interval (line segment) with end points $\mathbf{x}$ and $\mathbf{y}$, i.e. the set $\{\mathbf{z}: \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\,\mathbf{y},\ 0 \leqslant \alpha \leqslant 1\}$.

$(\mathbf{x}, \mathbf{y}) = \{\mathbf{z}: \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\,\mathbf{y},\ 0 < \alpha < 1\}$ is an open interval.

$(\mathbf{x}, \mathbf{y}] = \{\mathbf{z}: \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\,\mathbf{y},\ 0 < \alpha \leqslant 1\}$ is a half-open interval.

$[\mathbf{x}, \mathbf{y}) = \{\mathbf{z}: \mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\,\mathbf{y},\ 0 \leqslant \alpha < 1\}$ is a half-open interval.

$$\min\{\alpha, \beta\} = \begin{cases} \alpha & \text{if } \alpha \leqslant \beta, \\ \beta & \text{if } \alpha > \beta. \end{cases}$$

$\min\limits_{x \in X} \varphi(\mathbf{x})$ expresses the statement of the problem of minimizing a scalar function $\varphi(\mathbf{x})$ on a set X.

$\triangleq$ symbolizes definition and notation. For instance, $y \triangleq f(\mathbf{x}) + g(\mathbf{x})$ means that y denotes the sum $f(\mathbf{x}) + g(\mathbf{x})$.

$\varphi'(\mathbf{x})$ is the gradient of the function $\varphi(\mathbf{x})$ at the point $\mathbf{x}$: $\varphi'(\mathbf{x}) = \mathbf{grad}\ \varphi(x) = \nabla\varphi(x)$.

$$\operatorname{diam} X \underset{=}{\triangle} \sup_{\mathbf{x}', \mathbf{x}'' \in X} \| \mathbf{x}' - \mathbf{x}'' \|.$$

$$\operatorname{sgn} \alpha = \begin{cases} 1 & \text{if } \alpha \geqslant 0, \\ -1 & \text{if } \alpha < 0. \end{cases}$$

$P\{Q\}$ is the probability of the property Q to hold.

$P\{Q \mid R\}$ is the probability of the property Q to hold under the condition R.

$\triangle$ symbolizes the end of a proof.

$\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$ is a point where a minimum of the function $\varphi(\mathbf{x})$ on the set X is attained.

$X^* = \operatorname{Arg\,min} \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$ is the set of all points where the function $\varphi(\mathbf{x})$ attains minimum on the set X.

$E_n^+ = \{\mathbf{x} \in E_n, \ \mathbf{x} \geqslant \mathbf{0}\}$ is the set of all nonnegative points in the space E_n.

$:=$ is the symbol of assignment. For example, $\alpha := \beta$ means the assignment of the value β to α.

Chapter 2

ELEMENTS OF CONVEX ANALYSIS

2.1. Euclidean Space. Convex Sets

2.1.1. We shall deal with functions on sets belonging to a finite-dimensional Euclidean space E_n.

The set of all ordered systems $\mathbf{x}^T = (x_1, x_2, \ldots, x_n)$ of n real numbers $x_1, x_2, \ldots, x_n$ is called an n-dimensional Euclidean space if the following conditions are fulfilled. Let $\mathbf{x} \in E_n$, $\mathbf{y} \in E_n$ and let α be a real number; then

$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$ (addition),

$\alpha\mathbf{x} = (\alpha x_1, \alpha x_2, \ldots, \alpha x_n)$ (multiplication by a scalar),

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{n} x_i y_i \quad \text{(scalar product)}.$$

The elements $\mathbf{x}^T = (x_1, x_2, \ldots, x_n)$ are called *points* (*vectors*) and the numbers $x_1, x_2, \ldots, x_n$ are called their *coordinates*.

In a Euclidean space we introduce the notion of the *Euclidean norm* (the length of a vector)

$$\| \mathbf{x} \| = \langle \mathbf{x}, \mathbf{x} \rangle^{1/2}$$

possessing the following properties:

$$\| \mathbf{x} \| > 0$$

and $\| \mathbf{x} \| = 0$ if and only if $\mathbf{x} = \mathbf{0}$;

$$\| \alpha\mathbf{x} \| = | \alpha | \, \| \mathbf{x} \|;$$

$$\| \mathbf{x} + \mathbf{y} \| \leqslant \| \mathbf{x} \| + \| \mathbf{y} \|.$$

The Euclidean norm and the scalar product satisfy the Cauchy-Bunyakovskiĭ inequality:

$$\langle \mathbf{x}, \mathbf{y} \rangle \leqslant \| \mathbf{x} \| \cdot \| \mathbf{y} \|.$$

The Euclidean norm gives rise to convergence in E_n. A sequence $\{\mathbf{x}_m\}$ of points belonging to E_n is said to

converge to a point $\mathbf{x}$ for $m \to \infty$, i.e. $\lim\limits_{m\to\infty} \mathbf{x}_m = \mathbf{x}$, if

$$\lim_{m\to\infty} \| \mathbf{x}_m - \mathbf{x} \| = 0.$$

The point $\mathbf{x}$ is called the *limiting point* of the sequence.

We also give several more definitions.

2.1.2. The set

$$U_\varepsilon(\mathbf{x}) = \{\mathbf{y}: \| \mathbf{y} - \mathbf{x} \| \leqslant \varepsilon\}$$

will be called an ε-*neighborhood* of the point $\mathbf{x}$.

2.1.3. A set $X \subseteq E_n$ is said to be *closed* if it contains all its limiting points, i.e. points whose any neighborhood contains infinitely many points belonging to X.

2.1.4. A point $\mathbf{x} \in X$ is called an *interior point* of the set X if there is a neighborhood of x whose all points belong to X.

2.1.5. A point $\mathbf{x} \in X$ is called a *boundary point* of the set X if its any neighborhood contains points both belonging to the set X and not belonging to it. The set $G(X)$ consisting of all boundary points of the set X is called the *boundary* of the set X.

2.1.6. A set X in an n-dimensional Euclidean space E_n is said to be *convex* if, along with any two points $\mathbf{x} \in X$ and $\mathbf{y} \in X$, it contains the line segment $[\mathbf{x}, \mathbf{y}]$ joining them.

The convexity of a set X means that $\mathbf{x}, \mathbf{y} \in X$ implies $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\,\mathbf{y} \in X$ for all $0 \leqslant \alpha \leqslant 1$. For instance, in E_2 a line segment, a half-line, a whole straight line, a circle, a triangle, a half-plane, and the entire plane are examples of convex sets.

It can easily be seen that if a set X is determined by a system of linear equations and inequalities

$$X = \{\mathbf{x}: \; A\mathbf{x} \geqslant \mathbf{a}, \quad B\mathbf{x} = \mathbf{b}\},$$

where A and B are matrices, then it is convex and closed (the latter follows from the linearity and, consequently, the continuity of the operators A and B).

The reader can readily verify that the intersection of any number of convex sets is convex.

2.1.7. A point z is called a *convex combination* of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$ if $\mathbf{z} = \alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_m\mathbf{x}_m$, $\alpha_i \geqslant 0$ $(i = \overline{1, m})$, and $\alpha_1 + \alpha_2 + \ldots + \alpha_m = 1$.

2.1.8. *A convex set X contains all convex combinations of its points.*

Proof (by induction). According to the Definition 2.1.6 the set X contains the convex combination of any two of its points. Assume that X contains the convex combinations of any $m - 1$, $m > 1$, of its points. Consider a convex combination $\mathbf{z} = \alpha_1 \mathbf{x}_1 + \alpha_2 \mathbf{x}_2 + \ldots + \alpha_m \mathbf{x}_m$ of m arbitrary points belonging to X. It is obvious that there is at least one index i such that $\alpha_i < 1$. Without loss of generality, we can assume that $\alpha_1 < 1$. Then $\mathbf{y} = \beta_2 \mathbf{x}_2 + \ldots + \beta_m \mathbf{x}_m$, where $\beta_i = \alpha_i (1 - \alpha_1)^{-1}$, is a convex combination of the points $\mathbf{x}_2, \ldots, \mathbf{x}_m$, and, by the induction hypothesis, $\mathbf{y} \in X$. Since $\mathbf{z} = \alpha_1 \mathbf{x}_1 + (1 - \alpha_1)\, \mathbf{y}$ and $\alpha_1 \in [0, 1]$, by 2.1.6 the convexity of X implies that $\mathbf{z} \in X$. $\triangle$

2.2. Projection. Separation Theorems

2.2.1. By a *projection* of a point $\mathbf{v}$ on a set X is meant a point $\mathbf{p} \in X$ such that

$$\| \mathbf{p} - \mathbf{v} \| = \inf_{\mathbf{x} \in X} \| \mathbf{x} - \mathbf{v} \| \underset{=}{\triangle} \rho (\mathbf{v}, X) = \rho, \quad (2.1)$$

and $\rho = \rho (\mathbf{v}, X)$ is called the *distance* from the point $\mathbf{v}$ to the set X.

2.2.2. *For any closed set X and any point $\mathbf{v}$ there is a point $\mathbf{p} \in X$ which is a projection of $\mathbf{v}$ onto X. Moreover, if the set X is convex, then the point $\mathbf{p}$ is unique.*

Proof. If $\mathbf{v} \in X$, then obviously $\mathbf{p} = \mathbf{v}$ and $\rho = 0$. Let $\mathbf{v}$ be an exterior point relative to X: $\mathbf{v} \notin X$. By the definition of infimum, there is a sequence $\{\mathbf{x}_k\}$, $\mathbf{x}_k \in X$, such that

$$\lim_{k \to \infty} \| \mathbf{x}_k - \mathbf{v} \| = \rho.$$

Since the sequence $\{\mathbf{x}_k\}$ is bounded, there is a subsequence $\{\mathbf{x}_{k_i}\}$ such that

$$\lim_{i \to \infty} \mathbf{x}_{k_i} = \mathbf{p}.$$

Since the set X is closed, $\mathbf{p} \in X$. Finally, we have $\| \mathbf{p} - \mathbf{v} \| = \rho$.

To prove the uniqueness, we assume that there are two points $\mathbf{p}', \mathbf{p}'' \in X$, $\mathbf{p}' \neq \mathbf{p}''$, such that

$$\| \mathbf{p}' - \mathbf{v} \| = \| \mathbf{p}'' - \mathbf{v} \| = \rho.$$

Since the set X is convex, the point $\mathbf{z} = (1/2)(\mathbf{p}' + \mathbf{p}'')$ belongs to X. The points $\mathbf{v}$, $\mathbf{p}'$, $\mathbf{p}''$, and $\mathbf{z}$ lie in one plane and, since the triangle with vertex at $\mathbf{v}$, base $[\mathbf{p}', \mathbf{p}'']$, and height $[\mathbf{v}, \mathbf{z}]$ is isosceles, we have $\| \mathbf{z} - \mathbf{v} \| < \rho$, which contradicts the definition of ρ. $\triangle$

2.2.3. Theorem. *For a point $\mathbf{p} \in X$ to be the projection of a point $\mathbf{v}$ on a closed convex set X it is necessary and sufficient that for all $\mathbf{x} \in X$ the inequality*

$$\langle \mathbf{x} - \mathbf{p}, \ \mathbf{v} - \mathbf{p} \rangle \leqslant 0 \tag{2.2}$$

hold.

Proof. Let $\mathbf{p}$ be the projection of the point $\mathbf{v}$ on X. Take an arbitrary point $\mathbf{x} \in X$ and consider $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{p}$. By virtue of the convexity of X, the point $\mathbf{z}$ belongs to X for any $\alpha \in [0, 1]$. Since

$$\| \mathbf{z} - \mathbf{v} \|^2 = \alpha^2 \| \mathbf{x} - \mathbf{p} \|^2 + 2\alpha \langle \mathbf{x} - \mathbf{p}, \ \mathbf{p} - \mathbf{v} \rangle + \| \mathbf{p} - \mathbf{v} \|^2,$$

$$\| \mathbf{z} - \mathbf{v} \|^2 \geqslant \| \mathbf{p} - \mathbf{v} \|^2,$$

we have

$$\alpha^2 \| \mathbf{x} - \mathbf{p} \|^2 + 2\alpha \langle \mathbf{x} - \mathbf{p}, \ \mathbf{p} - \mathbf{v} \rangle \geqslant 0.$$

Since this inequality is fulfilled for all $\alpha \in [0, 1]$, we obtain

$$\langle \mathbf{x} - \mathbf{p}, \ \mathbf{p} - \mathbf{v} \rangle \geqslant 0,$$

which implies (2.2).

Now let (2.2) hold. Then for any $\mathbf{x} \in X$ we have

$$\| \mathbf{x} - \mathbf{v} \|^2 = \| (\mathbf{x} - \mathbf{p}) + (\mathbf{p} - \mathbf{v}) \|^2 = \| \mathbf{x} - \mathbf{p} \|^2 + 2 \langle \mathbf{x} - \mathbf{p}, \ \mathbf{p} - \mathbf{v} \rangle + \| \mathbf{p} - \mathbf{v} \|^2 \geqslant \| \mathbf{p} - \mathbf{v} \|^2,$$

i.e. $\mathbf{p}$ is the projection of $\mathbf{v}$ on X. $\triangle$

2.2.4. *For any $\mathbf{x} \in X$,*

$$\langle \mathbf{x} - \mathbf{v}, \ \mathbf{x} - \mathbf{p} \rangle = \langle \mathbf{x} - \mathbf{p}, \ \mathbf{x} - \mathbf{p} \rangle + \langle \mathbf{p} - \mathbf{v}, \ \mathbf{x} - \mathbf{p} \rangle \geqslant \| \mathbf{x} - \mathbf{p} \|^2,$$

$$\| \mathbf{x} - \mathbf{p} \| \leqslant \| \mathbf{x} - \mathbf{v} \|.$$

This follows from the fact that

$$\| \mathbf{x} - \mathbf{v} \|^2 = \| (\mathbf{x} - \mathbf{p}) + (\mathbf{p} - \mathbf{v}) \|^2$$
$$= \| \mathbf{x} - \mathbf{p} \|^2 + 2 \langle \mathbf{x} - \mathbf{p}, \mathbf{p} - \mathbf{v} \rangle$$
$$+ \| \mathbf{p} - \mathbf{v} \|^2 \geqslant \| \mathbf{x} - \mathbf{p} \|^2. \quad \triangle$$

2.2.5. By a *hyperplane* in E_n is meant a set having the form

$$\Pi = \{\mathbf{x}: \langle \mathbf{c}, \mathbf{x} \rangle = \lambda\},$$

where $\mathbf{c} \neq \mathbf{0}$. A hyperplane in E_n determines two half-spaces:

$$\{\mathbf{x}: \langle \mathbf{c}, \mathbf{x} \rangle \leqslant \lambda\} \text{ and } \{\mathbf{x}: \langle \mathbf{c}, \mathbf{x} \rangle \geqslant \lambda\}.$$

2.2.6. Separation theorem. *For any convex and closed set X and any point* $\mathbf{v}$ *not belonging to X there is a hyperplane* Π *such that*

$$\langle \mathbf{c}, \mathbf{v} \rangle = \lambda, \tag{2.3}$$

and, for all $\mathbf{x} \in X$,

$$\langle \mathbf{c}, \mathbf{x} \rangle < \lambda. \tag{2.4}$$

Proof. Let $\mathbf{p}$ be the projection of the point $\mathbf{v}$ on X. Consider the hyperplane

$$\Pi = \{\mathbf{x}: \langle \mathbf{c}, \mathbf{x} \rangle = \lambda, \quad \mathbf{c} = \mathbf{v} - \mathbf{p}, \quad \lambda = \langle \mathbf{c}, \mathbf{v} \rangle\},$$

for which (2.3) holds. The inequality (2.2) implies

$$\langle \mathbf{x}, \mathbf{v} - \mathbf{p} \rangle \leqslant \langle \mathbf{p}, \mathbf{v} - \mathbf{p} \rangle < \langle \mathbf{v}, \mathbf{v} - \mathbf{p} \rangle, \quad \mathbf{x} \in X.$$

The right-hand inequality follows from (2.1) and from the fact that $\rho > 0$. Finally,

$$\langle \mathbf{c}, \mathbf{x} \rangle = \langle \mathbf{v} - \mathbf{p}, \mathbf{x} \rangle < \langle \mathbf{v} - \mathbf{p}, \mathbf{v} \rangle = \langle \mathbf{c}, \mathbf{v} \rangle = \lambda,$$

i.e. (2.4) is true. $\triangle$

Remark. The theorem has an obvious geometrical meaning: there is a hyperplane Π passing through the point $\mathbf{v}$ such that X belongs to one of the half-spaces determined by Π (Fig. 2.1).

2.2.7. Theorem on the supporting hyperplane. *At any boundary point* $\mathbf{x}^0$ *of a convex set X there exists a supporting hyperplane, i.e. there are* $\mathbf{c} \neq \mathbf{0}$ *and* λ *such that* $\Pi = \{\mathbf{x}: \langle \mathbf{c}, \mathbf{x} \rangle = \lambda\}$, $\lambda = \langle \mathbf{c}, \mathbf{x}^0 \rangle$, *and* $\langle \mathbf{c}, \mathbf{x} \rangle \leqslant \lambda$ *for all* $\mathbf{x} \in X$.

Proof. Consider a sequence of points $\{\mathbf{v}_k\}$ exterior relative to $\overline{X}$ ($\overline{X}$ is the closure of X) and such that

$$\lim_{k\to\infty} \mathbf{v}_k = \mathbf{x}^0.$$

By Theorem 2.2.6, for each $\mathbf{v}_k$ there is

$$\Pi_k = \{\mathbf{x}: \langle \mathbf{c}_k, \mathbf{x}\rangle = \lambda_k\},$$

where $\lambda_k = \langle \mathbf{c}_k, \mathbf{v}_k\rangle$ and $\langle \mathbf{c}_k, \mathbf{x}\rangle < \lambda_k$ for all $\mathbf{x} \in \overline{X}$.

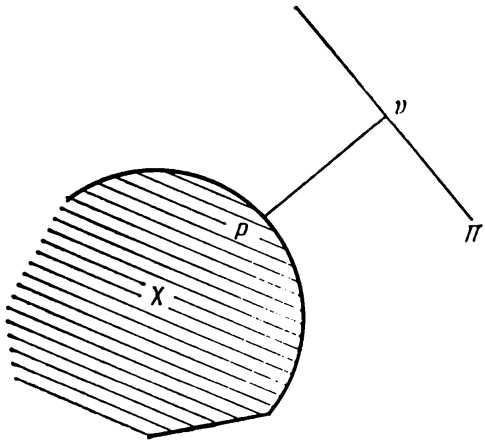


Fig. 2.1

Without loss of generality, we can assume that $\|\mathbf{c}_k\| = 1$. We do not change the notation and put

$$\lim_{k\to\infty} \mathbf{c}_k = \mathbf{c}.$$

Passing to the limit in the relations determining Π_k we find

$$\langle \mathbf{c}, \mathbf{x}^0\rangle = \lim_{k\to\infty} \langle \mathbf{c}_k, \mathbf{v}_k\rangle \triangleq \lambda,$$

and $\langle \mathbf{c}, \mathbf{x}\rangle \leqslant \lambda$ for all $\mathbf{x} \in X$.

Thus,

$$\Pi = \{\mathbf{x}: \langle \mathbf{c}, \mathbf{x}\rangle = \lambda\}$$

is a supporting hyperlane. $\triangle$

Remark. It can easily be seen that if there is a tangent hyperplane at the point $\mathbf{x}^0$, then it coincides with the supporting hyperplane (Fig. 2.2), and in this case the supporting hyperplane is unique. However, the notion

of a supporting hyperplane is much more general than that of a tangent hyperplane. Figure 2.3 demonstrates the case where at a point $\mathbf{x}^0$ there is no tangent whereas there are supporting straight lines at it, and any vector lying "between" $\mathbf{c}_1$ and $\mathbf{c}_2$ can be taken as $\mathbf{c}$.

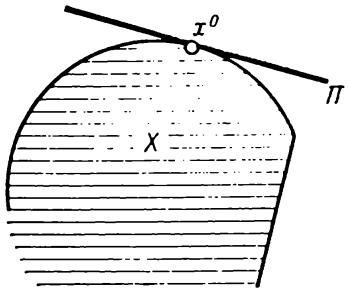


Fig. 2.2

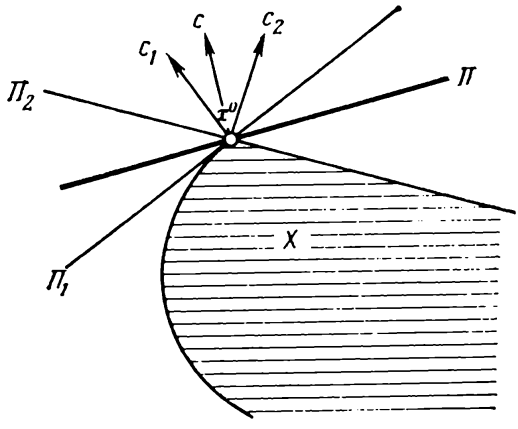


Fig. 2.3

2.2.8. Theorem on the separating hyperplane. *If the set X_0 of interior points of a convex set X is nonempty and does not intersect a convex set Y ($X_0 \cap Y = \varnothing$), then for the sets X and Y there exists a separating hyperplane* Π, *i.e. there is a vector* $\mathbf{c} \neq \mathbf{0}$ *such that*

$$\langle \mathbf{c}, \mathbf{y} \rangle \leqslant \langle \mathbf{c}, \mathbf{x} \rangle$$

for all $\mathbf{y} \in Y$ *and* $\mathbf{x} \in X$.

Proof. The set

$$Z = \{\mathbf{z}: \mathbf{z} = \mathbf{y} - \mathbf{x}, \mathbf{y} \in Y, \mathbf{x} \in X_0\}$$

is convex, and $\mathbf{z} = \mathbf{0}$ is not its interior point. Therefore Theorems 2.2.6 and 2.2.7 imply the existence of $\mathbf{c} \neq \mathbf{0}$ such that

$$\langle \mathbf{c}, \mathbf{z} \rangle = \langle \mathbf{c}, \mathbf{y} - \mathbf{x} \rangle \leqslant \langle \mathbf{c}, \mathbf{0} \rangle = 0$$

for all $\mathbf{y} \in Y$ and $\mathbf{x} \in X_0$. This inequality remains valid for all $\mathbf{y} \in Y$ and $\mathbf{x} \in X$ as well since the passage to the limit does not violate slack inequalities. $\triangle$

Remark. It should be noted that the requirement that the set X_0 be nonempty is essential here since the statement of the theorem involves the interior of the set X relative to the space E_n. For instance, it is obvious that

in the three-dimensional space the axis z and the plane $z = 0$ are not separable in the above sense although they have no common interior points.

2.2.9. Definition. A point $\mathbf{x}$ belonging to a set X is called an *extreme point* if X contains no points $\mathbf{x}'$ and $\mathbf{x}''$, $\mathbf{x}' \neq \mathbf{x}''$, such that

$$\mathbf{x} = \alpha\mathbf{x}' + (1 - \alpha)\,\mathbf{x}''$$

for some $\alpha \in (0, 1)$.

For example, for a disk any point of its bounding circle is an extreme point. All the vertices of a convex polyhedron are extreme points.

2.2.10. Representation theorem. *Any point* $\mathbf{x}^0$ *belonging to a convex, closed, and bounded set* X *can be represented as a convex combination of a finite number of extreme points of the set.*

Proof (by induction on the lowest dimension n of the space E_n containing the set X).

If $n = 1$, then X is a line segment, and the assertion of the theorem is obvious.

Assume that the theorem is true for $n = k - 1$. Let $X \subset E_k$. We consider the following two cases.

(1) $\mathbf{x}^0$ is a boundary point of X. We construct at this point a supporting hyperplane to X:

$$\pi = \{\mathbf{x}\colon \langle \mathbf{c}, \mathbf{x}\rangle = \langle \mathbf{c}, \mathbf{x}^0\rangle\}.$$

The set $X_0 = X \cap \pi$ is the intersection of the convex, closed, and bounded set X with the convex and closed set π and therefore it is itself convex, closed, and bounded, and, moreover, there is a $(k - 1)$-dimensional subspace containing X_0 (since $X_0 \subset \pi$). By the induction hypothesis, for $\mathbf{x}^0 \in X_0$ there are extreme points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ of the set X_0 such that

$$\mathbf{x}^0 = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \quad \alpha_i \geqslant 0, \quad i = \overline{1, N},$$

$$\sum_{i=1}^{N} \alpha_i = 1.$$

We shall show that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ are extreme points for X as well. Assume the contrary, i.e. let for a point $\mathbf{x}_i$

there be $\mathbf{x}', \mathbf{x}'' \in X$, $\mathbf{x}' \neq \mathbf{x}''$, and $\alpha \in (0, 1)$ such that

$$\mathbf{x}_i = \alpha \mathbf{x}' + (1 - \alpha)\, \mathbf{x}''.$$

Since $\mathbf{x}_i \in X_0 \subset \pi$, we have

$$\langle \mathbf{c}, \mathbf{x}_i \rangle = \langle \mathbf{c}, \mathbf{x}^0 \rangle,$$

and, since π is a supporting hyperplane to X,

$$\langle \mathbf{c}, \mathbf{x}' \rangle \leqslant \langle \mathbf{c}, \mathbf{x}^0 \rangle, \quad \langle \mathbf{c}, \mathbf{x}'' \rangle \leqslant \langle \mathbf{c}, \mathbf{x}^0 \rangle.$$

Moreover, $0 < \alpha < 1$ implies

$$\langle \mathbf{c}, \mathbf{x}' \rangle = \frac{1}{\alpha} [\langle \mathbf{c}, \mathbf{x}_i \rangle - (1 - \alpha) \langle \mathbf{c}, \mathbf{x}'' \rangle]$$

$$\geqslant \frac{1}{\alpha} [\langle \mathbf{c}, \mathbf{x}^0 \rangle - (1 - \alpha) \langle \mathbf{c}, \mathbf{x}^0 \rangle] = \langle \mathbf{c}, \mathbf{x}^0 \rangle.$$

The last two conditions show that $\mathbf{x}' \in \pi$ (since $\langle \mathbf{c}, \mathbf{x}' \rangle = \langle \mathbf{c}, \mathbf{x}^0 \rangle$); however, since $\mathbf{x}' \in X$, it follows that $\mathbf{x}' \in X_0 = X \cap \pi$. It is similarly proved that $\mathbf{x}'' \in X_0$. Therefore our assumption contradicts the fact that $\mathbf{x}_i$ is an extreme point of the set X_0.

(2) Now let $\mathbf{x}^0$ be an interior point of the set X. Consider a straight line l passing through $\mathbf{x}^0$. The intersection $l \cap X$ is a line segment with end points $\bar{\mathbf{x}}$ and $\bar{\bar{\mathbf{x}}}$ belonging to the boundary of the set X, and, since $\mathbf{x}^0$ is an interior point of X, there is $\alpha \in (0, 1)$ such that

$$\mathbf{x}^0 = \alpha \bar{\mathbf{x}} + (1 - \alpha)\, \bar{\bar{\mathbf{x}}}.$$

Since the theorem is true for the boundary points $\bar{\mathbf{x}}$ and $\bar{\bar{\mathbf{x}}}$, it is true for $\mathbf{x}^0$ as well. Indeed, for the boundary points we have

$$\bar{\mathbf{x}} = \sum_{i=1}^{N_1} \beta_i \mathbf{y}_i, \quad \sum_{i=1}^{N_1} \beta_i = 1, \quad \beta_i \geqslant 0, \quad 1 = \overline{1, N_1},$$

$$\bar{\bar{\mathbf{x}}} = \sum_{i=1}^{N_2} \gamma_i \mathbf{z}_i, \quad \sum_{i=1}^{N_2} \gamma_i = 1, \quad \gamma_i \geqslant 0, \quad i = \overline{1, N_2},$$

where $\mathbf{y}_i$ and $\mathbf{z}_i$ are extreme points of the set X. Therefore,

$$\mathbf{x}^0 = \sum_{i=1}^{N_1} \alpha \beta_i \mathbf{y}_i + \sum_{i=1}^{N_2} (1 - \alpha)\, \gamma_i \mathbf{z}_i,$$

which implies the assertion of the theorem. $\triangle$

2.3. Cone. Farkas Theorem

In the theory of mathematical programming an important role is played by the Farkas theorem that will be proved in the end of this section. Before this we shall establish some auxiliary notions and facts which, however, are themselves of interest.

2.3.1. A set K is called a *cone* if $\mathbf{x} \in K$ implies $\lambda\mathbf{x} \in K$ for all $\lambda > 0$. For instance, the whole space E_n, as well as its any subspace, is a cone. The nonnegative orthant $\{\mathbf{x}: \mathbf{x} \geqslant \mathbf{0}\}$ is also a cone. It is obvious that the sets

$$\{\mathbf{x}: A\mathbf{x} \leqslant \mathbf{0}\} \quad \text{and} \quad \{\mathbf{y}: \mathbf{y} = A\mathbf{x}, \mathbf{x} \geqslant \mathbf{0}\}$$

are cones.

The following fact will be used when proving the Farkas theorem.

2.3.2. *The set*

$$Y = \{\mathbf{y}: \mathbf{y} = A\mathbf{x}, \mathbf{x} \geqslant \mathbf{0}\}$$

is closed.

Proof. Let $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$. We shall prove the assertion by induction on the number m. For $m = 1$ the set Y is a half-line and hence it is closed. Assume that for $m = k - 1$ the cone $\bar{Y}$ generated by the vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{k-1}$ is closed.

(1) If the cone Y contains the vectors $-\mathbf{a}_1, -\mathbf{a}_2, \ldots, -\mathbf{a}_k$, then it is a subspace of a dimension not exceeding k, and, consequently, it is a closed set.

(2) Assume that at least one of the vectors $-\mathbf{a}_1, -\mathbf{a}_2, \ldots, -\mathbf{a}_k$ does not belong to Y, say, $-\mathbf{a}_k \notin Y$. Each $\mathbf{y} \in Y$ can be represented as $\mathbf{y} = \bar{\mathbf{y}} + \alpha\mathbf{a}_k$, $\alpha \geqslant 0$, where $\bar{\mathbf{y}} \in \bar{Y}$. Consider a sequence $\{\mathbf{y}_n\}$ converging to $\mathbf{y}$. By what has been said, $\mathbf{y}_n = \bar{\mathbf{y}}_n + \alpha_n\mathbf{a}_k$, $\alpha_n \geqslant 0$, for all indices n. If the sequence $\{\alpha_n\}$ is bounded, we can assume, without loss of generality, that $\lim\limits_{n\to\infty} \alpha_n = \alpha$, and, consequently, $\mathbf{y} - \alpha\mathbf{a}_k = \lim\limits_{n\to\infty} (\mathbf{y}_n - \alpha_n\mathbf{a}_k) = \lim\limits_{n\to\infty} \bar{\mathbf{y}}_n \underset{=}{\triangle} \bar{\mathbf{y}} \in \bar{Y}$ since $\bar{Y}$ is closed. Therefore $\mathbf{y} = \bar{\mathbf{y}} + \alpha\mathbf{a}_k \in Y$. Thus, if the sequence $\{\alpha_n\}$ is bounded, then the set Y is closed.

We now assume that $\alpha_n \to +\infty$ for $n \to \infty$. Then $(\alpha_n)^{-1}\mathbf{y}_n = (\alpha_n)^{-1}\bar{\mathbf{y}}_n + \mathbf{a}_k$. Since $\lim\limits_{n\to\infty} (\alpha_n)^{-1}\mathbf{y}_n = \mathbf{0}$,

we have $\bar{\mathbf{y}} \triangleq \lim\limits_{n\to\infty} (\alpha_n)^{-1} \bar{\mathbf{y}}_n = -\mathbf{a}_k$. However, since $\bar{Y}$ is closed, we have $\bar{\mathbf{y}} \in \bar{Y}$, and, consequently, $-\mathbf{a}_k \in \bar{Y}$, which is a contradiction. $\triangle$

2.3.3. Let B be an $m \times n$ matrix and let $\mathbf{v} \in E_n$.

The Farkas theorem. *The inequality* $\langle \mathbf{v}, \mathbf{x} \rangle \leqslant 0$ *holds for all* $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \leqslant \mathbf{0}\}$ *if and only if there is a vector* $\mathbf{u} \geqslant \mathbf{0}$ *such that* $\mathbf{v} = B^T\mathbf{u}$.

Proof. Sufficiency. Let the relations $\mathbf{u} \geqslant \mathbf{0}$ and $\mathbf{v} = B^T\mathbf{u}$ hold. Then for any $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \leqslant \mathbf{0}\}$ we have $\langle \mathbf{v}, \mathbf{x} \rangle = \langle B^T\mathbf{u}, \mathbf{x} \rangle = \langle \mathbf{u}, B\mathbf{x} \rangle \leqslant 0$.

Necessity. Let $\langle \mathbf{v}, \mathbf{x} \rangle \leqslant 0$ for all $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \leqslant \mathbf{0}\}$. Consider the cone

$$Y = \{\mathbf{y}: \mathbf{y} = B^T\mathbf{u}, \quad \mathbf{u} \geqslant \mathbf{0}\}.$$

If $\mathbf{v} \in Y$, the theorem is proved. Assume that $\mathbf{v} \notin Y$. The set Y is convex (see 2.1.6) and closed (see 2.3.2), and therefore, by Theorem 2.2.6, there is a vector $\mathbf{c} \neq \mathbf{0}$ such that

$$\langle \mathbf{c}, \mathbf{y} \rangle < \langle \mathbf{c}, \mathbf{v} \rangle \tag{2.5}$$

for all $\mathbf{y} \in Y$.

Since $\lambda\mathbf{y} \in Y$ for all $\lambda \geqslant 0$, we conclude from (2.5) that $\lambda \langle \mathbf{c}, \mathbf{y} \rangle < \langle \mathbf{c}, \mathbf{v} \rangle$ for all $\lambda \geqslant 0$. Hence, $\langle \mathbf{c}, \mathbf{y} \rangle \leqslant 0$. However,

$$\langle \mathbf{c}, \mathbf{y} \rangle = \langle \mathbf{c}, B^T\mathbf{u} \rangle = \langle \mathbf{u}, B\mathbf{c} \rangle \leqslant 0.$$

Since this holds for all $\mathbf{u} \geqslant \mathbf{0}$,

$$B\mathbf{c} \leqslant \mathbf{0}. \tag{2.6}$$

Now, since $\mathbf{y} = \mathbf{0} \in Y$, (2.5) implies

$$\langle \mathbf{c}, \mathbf{v} \rangle > 0. \tag{2.7}$$

Taking $\mathbf{x} = \mathbf{c}$ we obtain from (2.6) and (2.7) a contradiction to the conditions of the theorem. $\triangle$

It will be of use to give a geometrical interpretation of the Farkas theorem. Let

$$B = \begin{bmatrix} \mathbf{b}_1^T \\ \mathbf{b}_2^T \\ \mathbf{b}_3^T \end{bmatrix}$$

and let $K = \{\mathbf{x}: B\mathbf{x} \leqslant \mathbf{0}\}$.

The cone K is the set of all vectors $\mathbf{x}$ forming non-acute angles with each of the vectors $\mathbf{b}_1$, $\mathbf{b}_2$, and $\mathbf{b}_3$ (see Fig. 2.4 where the cones K and $Y = \{\mathbf{y}: \mathbf{y} = B^T\mathbf{u}, \mathbf{u} \geqslant \mathbf{0}\}$ are shaded with vertical and horizontal lines, respectively).

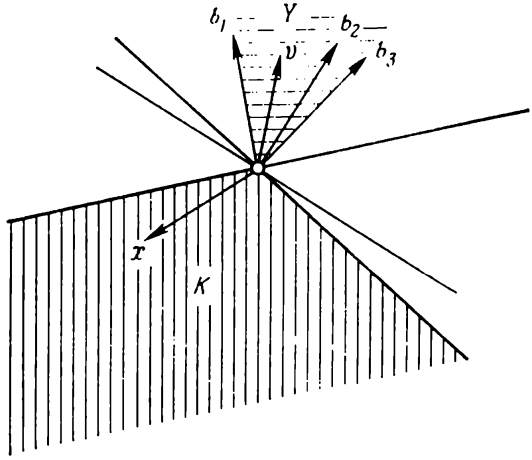


Fig. 2.4

The geometrical meaning of the theorem now becomes clear: for the angle between $\mathbf{v}$ and any $\mathbf{x} \in K$ to be non-acute it is necessary and sufficient that $\mathbf{v}$ belong to the cone Y.

2.3.4. Corollary. *For any matrix B and any vector $\mathbf{v}$ the following alternative holds: either the system*

$$B\mathbf{x} \geqslant \mathbf{0}, \quad \langle \mathbf{v}, \mathbf{x} \rangle < 0, \tag{2.8}$$

or the system

$$\mathbf{v} = B^T\mathbf{u}, \quad \mathbf{u} \geqslant \mathbf{0} \tag{2.9}$$

possesses a solution.

Proof. If (2.9) holds, then the Farkas theorem implies that for any $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \geqslant \mathbf{0}\}$ we have $\langle \mathbf{v}, \mathbf{x} \rangle \geqslant 0$, and, consequently, the system (2.8) is unsolvable. Assume now that the system (2.9) is unsolvable. Then, by the Farkas theorem, the condition $\langle \mathbf{v}, \mathbf{x} \rangle \geqslant 0$ is violated for all $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \geqslant \mathbf{0}\}$, and hence $\langle \mathbf{v}, \mathbf{x} \rangle < 0$ for some $\mathbf{x} \in \{\mathbf{x}: B\mathbf{x} \geqslant \mathbf{0}\}$. $\triangle$

2.3.5. The above corollary admits of a simple modification which is of use for a number of applications.

Corollary. *For any matrix B and any vector* $\mathbf{v}$ *the following alternative takes place: either the system*

$$B\mathbf{x} \geqslant \mathbf{0}, \quad \mathbf{x} \geqslant \mathbf{0}, \quad \langle \mathbf{v}, \mathbf{x} \rangle < 0 \tag{2.10}$$

or the system

$$\mathbf{v} \geqslant B^T\mathbf{u}, \quad \mathbf{u} \geqslant \mathbf{0} \tag{2.11}$$

possesses a solution.

Proof. Write the system (2.10) in the form (2.8):

$$\bar{B}\mathbf{x} \geqslant \mathbf{0}, \quad \langle \mathbf{v}, \mathbf{x} \rangle < 0.$$

Here $\bar{B} = \begin{bmatrix} B \\ E \end{bmatrix}$ where E is the unit matrix. The assertion of the corollary obviously follows from 2.3.4 in application to the matrix $\bar{B}$. $\triangle$

2.3.6. Representation theorem. Consider a cone $S = \{\mathbf{x}: B\mathbf{x} = \mathbf{0}, \mathbf{x} \geqslant \mathbf{0}\}$. A vector $\mathbf{x}$ belonging to the cone S will be called an *edge* if in S there are no $\mathbf{x}'$, $\mathbf{x}''$, $\mathbf{x}' \neq \mathbf{x}''$, such that for some $\lambda_1 > 0$ and $\lambda_2 > 0$ the equation $\mathbf{x} = \lambda_1\mathbf{x}' + \lambda_2\mathbf{x}''$ holds.

We now consider the hyperplane

$$L = \left\{\mathbf{x}: \sum_{i=1}^{n} x_i = 1, \quad x_i \geqslant 0, \ i = \overline{1, n}\right\}.$$

It is obvious that the set $Q = S \cap L$ is convex and bounded.

If $\mathbf{x}_1^*, \mathbf{x}_2^*, \ldots, \mathbf{x}_k^*$ are extreme points of the set Q, then the vectors $\alpha_1\mathbf{x}_1^*, \alpha_2\mathbf{x}_2^*, \ldots, \alpha_k\mathbf{x}_k^*$, $\alpha_i \geqslant 0$ $(i = \overline{1, k})$ are edges of the cone S. Indeed, if $\mathbf{z} = \alpha_j\mathbf{x}_j^*$ is not an edge, then there are $\mathbf{x}' \in S$ and $\mathbf{x}'' \in S$ such that $\mathbf{z} = \lambda_1\mathbf{x}' + \lambda_2\mathbf{x}''$ for some $\lambda_1, \lambda_2 > 0$. It follows from the definition of the set Q that there are $\beta_1, \beta_2 > 0$ such that $\mathbf{y}' = \beta_1\mathbf{x}' \in Q$, $\mathbf{y}'' = \beta_2\mathbf{x}'' \in Q$. However,

$$\mathbf{x}_j^* = \frac{1}{\alpha_j}\mathbf{z} = \gamma_1\mathbf{y}' + \gamma_2\mathbf{y}'', \quad \gamma_1 = \frac{\lambda_1}{\alpha_j\beta_1}, \quad \gamma_2 = \frac{\lambda_2}{\alpha_j\beta_2},$$

$$\sum_{i=1}^{k} \frac{1}{\alpha_j} z_i = 1, \quad \sum_{i=1}^{k} y_i' = 1, \quad \sum_{i=1}^{k} y_i'' = 1,$$

and therefore $\gamma_1 + \gamma_2 = 1$, which contradicts the assumption that $\mathbf{x}_j^*$ is an extreme point of the set Q.

The above-proved assertion and Theorem 2.2.10 make it possible to state the representation theorem:

Any vector $\mathbf{x}$ *belonging to the cone* S *can be represented as a positive combination of its edges, i.e. there are numbers* $\lambda_1 \geqslant 0,\ \lambda_2 \geqslant 0,\ \ldots,\ \lambda_k \geqslant 0$ *such that* $\mathbf{x} = \sum_{i=1}^{k} \lambda_i \mathbf{x}_i^*$.

2.4. Convex Functions

2.4.1. Definition. A function $\varphi(\mathbf{x})$ on a convex set X is said to be *convex* if for any $\mathbf{x}, \mathbf{y} \in X$ and all $\alpha \in [0, 1]$ the inequality

$$\varphi(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \leqslant \alpha\varphi(\mathbf{x}) + (1 - \alpha)\varphi(\mathbf{y}) \quad (2.12)$$

holds. Figure 2.5 represents a convex function. It is

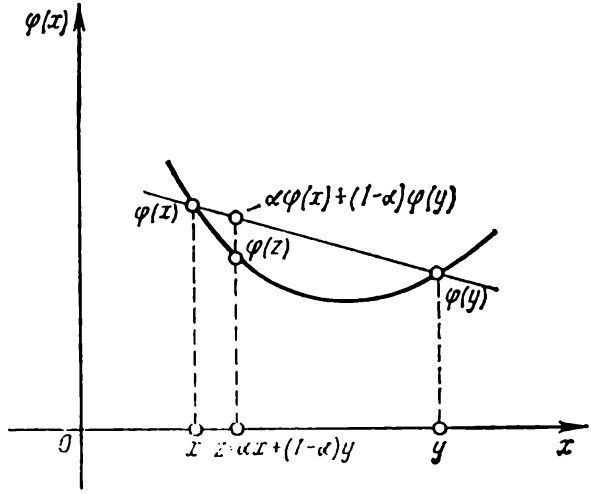


Fig. 2.5

obvious that each point of any chord of the graph of the function φ either lies above the graph or belongs to it.

2.4.2. By a *concave function* is meant a function $\varphi(\mathbf{x})$ for which the function $-\varphi(\mathbf{x})$ is convex. Thus, if

$$\varphi(\alpha\mathbf{x} + (1 - \alpha)\mathbf{y}) \geqslant \alpha\varphi(\mathbf{x}) + (1 - \alpha)\varphi(\mathbf{y}) \quad (2.13)$$

for any $\mathbf{x}, \mathbf{y} \in X$ and all $\alpha \in [0, 1]$, then $\varphi(\mathbf{x})$ is a concave function.

2.4.3. If (2.12) is a strict inequality for any $\alpha \in (0, 1)$, the function $\varphi(\mathbf{x})$ is called *strictly convex.*

2.4.4. An example of a convex function is a quadratic function with a positive definite matrix.

Since this property of quadratic functions will be used in what follows, its proof is itself of interest and we present it here.

For a quadratic function

$$\varphi(\mathbf{x}) = \langle \mathbf{x}, B\mathbf{x} \rangle + \langle \mathbf{p}, \mathbf{x} \rangle$$

to be convex it is necessary and sufficient that the symmetric matrix B be positive definite.

Proof.

$$\begin{aligned}\varphi(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) &= \alpha^2 \langle \mathbf{x}, B\mathbf{x} \rangle + 2\alpha(1-\alpha)\langle \mathbf{x}, B\mathbf{y} \rangle \\ &\quad + (1-\alpha)^2 \langle \mathbf{y}, B\mathbf{y} \rangle + \alpha \langle \mathbf{p}, \mathbf{x} \rangle + (1-\alpha)\langle \mathbf{p}, \mathbf{y} \rangle \\ &= \alpha\varphi(\mathbf{x}) + (1-\alpha)\varphi(\mathbf{y}) - \alpha(1-\alpha) \\ &\quad \times \langle \mathbf{x}-\mathbf{y}, B(\mathbf{x}-\mathbf{y}) \rangle,\end{aligned}$$

and for $\alpha \in (0, 1)$ we have

$$\alpha(1-\alpha)\langle \mathbf{x}-\mathbf{y}, B(\mathbf{x}-\mathbf{y}) \rangle \geqslant 0$$

if and only if B is positive definite. This and the definition of convexity imply the assertion of the theorem. △

Remark. It is obvious that for a quadratic function $\varphi(\mathbf{x})$ to be strictly convex it is necessary and sufficient that the matrix B be strictly positive definite.

We consider a number of properties of convex functions.

2.4.5. *For any convex function $\varphi(\mathbf{x})$ on a convex set X and any number λ the set $Z = \{\mathbf{x} \in X: \varphi(\mathbf{x}) \leqslant \lambda\}$ is convex.*

Proof. According to the definition 2.1.6, it suffices to prove that $\mathbf{x}, \mathbf{y} \in Z$ implies $\mathbf{z} = \alpha\mathbf{x} + (1-\alpha)\mathbf{y} \in Z$ for all $\alpha \in [0, 1]$. It follows from the convexity of the set X that $\mathbf{z} \in X$. Using the inequality (2.12) we find

$$\begin{aligned}\varphi(\mathbf{z}) = \varphi(\alpha\mathbf{x} + (1-\alpha)\mathbf{y} &\leqslant \alpha\varphi(\mathbf{x}) \\ &\quad + (1-\alpha)\varphi(\mathbf{y}) \leqslant \alpha\lambda + (1-\alpha)\lambda = \lambda. \quad \triangle\end{aligned}$$

Remark. It is obvious that if $\varphi(\mathbf{x})$ is a concave function, then the set $Z = \{\mathbf{x} \in X: \varphi(\mathbf{x}) \geqslant \lambda\}$ is convex.

2.4.6. Jensen's inequality. *If* $\varphi(\mathbf{x})$ *is a convex function on a convex set* X *and*

$$\mathbf{z} = \sum_{i=1}^{m} \alpha_i \mathbf{x}_i, \quad \sum_{i=1}^{m} \alpha_i = 1,$$

$$\alpha_i \geqslant 0, \quad \mathbf{x}_i \in X, \quad i = \overline{1, m},$$

then

$$\varphi\left(\sum_{i=1}^{m} \alpha_i \mathbf{x}_i\right) \leqslant \sum_{i=1}^{m} \alpha_i \varphi(\mathbf{x}_i). \tag{2.14}$$

Proof (by induction). The inequality (2.14) obviously holds for $m = 1$. Assume that (2.14) is valid for $m - 1$, $m > 1$. It follows from (2.1.8) that $\mathbf{z} \in X$. If $\alpha_m = 1$, then $\alpha_1 = \alpha_2 = \ldots = \alpha_{m-1} = 0$, and (2.14) involves the sign of equality. If $0 \leqslant \alpha_m < 1$, then the convexity of φ and the induction hypothesis imply

$$\varphi(\mathbf{z}) = \varphi\left(\sum_{i=1}^{m} \alpha_i \mathbf{x}_i\right) = \varphi\left((1-\alpha_m) \sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} \mathbf{x}_i + \alpha_m \mathbf{x}_m\right)$$

$$\leqslant (1-\alpha_m)\, \varphi\left(\sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} \mathbf{x}_i\right) \alpha_m \varphi(\mathbf{x}_m)$$

$$\leqslant (1-\alpha_m) \sum_{i=1}^{m-1} \frac{\alpha_i}{1-\alpha_m} \varphi(\mathbf{x}_i) + \alpha_m \varphi(\mathbf{x}_m) = \sum_{i=1}^{m} \alpha_i \varphi(\mathbf{x}_i). \quad \triangle$$

2.4.7. We present (without proof) the following important property of convex functions.

A convex function $\varphi(\mathbf{x})$ *on a convex set* X *is continuous at each interior point of the set and possesses the derivative in any direction* $\mathbf{s}$ ($\| \mathbf{s} \| = 1$) *at it*:

$$\frac{\partial \varphi(\mathbf{x})}{\partial \mathbf{s}} = \lim_{\lambda \to +0} \frac{\varphi(\mathbf{x} + \lambda \mathbf{s}) - \varphi(\mathbf{x})}{\lambda}.$$

2.4.8. The following two properties of convex functions are applied in the penalty function method.

If $\varkappa(\mathbf{x})$ *is a convex function on a convex set* X, *then the function*

$$\varphi(\mathbf{x}) = \max\{\varkappa(\mathbf{x}), 0\}$$

is also convex on X.

Indeed, for any $\mathbf{x}, \mathbf{y} \in X$,

$$\varphi(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) = \max\{\varkappa(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}), 0\}$$
$$\leqslant \max\{\alpha\varkappa(\mathbf{x}) + (1-\alpha)\varkappa(\mathbf{y}), 0\} \leqslant \alpha \max\{\varkappa(\mathbf{x}), 0\}$$
$$+ (1-\alpha)\max\{\varkappa(\mathbf{y}), 0\} = \alpha\varphi(\mathbf{x}) + (1-\alpha)\varphi(\mathbf{y}). \quad \triangle$$

If $\varphi(\mathbf{x})$ is convex and nonnegative on a convex set X, then $\varphi^2(\mathbf{x})$ is also convex on X.

Indeed, by the convexity and nonnegativity of $\varphi(\mathbf{x})$, we have

$$\varphi^2(\alpha\mathbf{x} + (1-\alpha)\mathbf{y})$$
$$\leqslant \alpha^2\varphi^2(\mathbf{x}) + 2\alpha(1-\alpha)\varphi(\mathbf{x})\varphi(\mathbf{y}) + (1-\alpha)^2\varphi^2(\mathbf{y})$$
$$= \alpha\varphi^2(\mathbf{x}) + (1-\alpha)\varphi^2(\mathbf{y}) - \alpha(1-\alpha)(\varphi(\mathbf{x}) - \varphi(\mathbf{y}))^2$$
$$\leqslant \alpha\varphi^2(\mathbf{x}) + (1-\alpha)\varphi^2(\mathbf{y}). \quad \triangle$$

2.4.9. The proposition below establishes an important property of differentiable convex functions which will often be used in what follows.

A function $\varphi(\mathbf{x})$ differentiable on a convex set X is convex or concave if and only if, for any $\mathbf{x} \in X$ and $\mathbf{y} \in Y$,

$$\langle \varphi'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leqslant \varphi(\mathbf{y}) - \varphi(\mathbf{x}) \qquad (2.15)$$

or

$$\langle \varphi'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \geqslant \varphi(\mathbf{y}) - \varphi(\mathbf{x}),$$

respectively.

Proof. Let $\varphi(\mathbf{x})$ be convex. Then for any $\mathbf{x}, \mathbf{y} \in X$, $\mathbf{x} \neq \mathbf{y}$, and all α such that $0 < \alpha \leqslant 1$ we have the inequality

$$\varphi(\mathbf{x} + \alpha(\mathbf{y} - \mathbf{x})) \leqslant \varphi(\mathbf{x}) + \alpha[\varphi(\mathbf{y}) - \varphi(\mathbf{x})],$$

whence

$$\|\mathbf{y} - \mathbf{x}\| \frac{\varphi(\mathbf{x} + \beta\mathbf{s}) - \varphi(\mathbf{x})}{\beta} \leqslant \varphi(\mathbf{y}) - \varphi(\mathbf{x}),$$

where $\mathbf{s} = (\mathbf{y} - \mathbf{x})\|\mathbf{y} - \mathbf{x}\|^{-1}$ and $\beta = \alpha\|\mathbf{y} - \mathbf{x}\|$.

Passing to the limit for $\beta \to 0$ in the last inequality we obtain

$$\frac{\partial\varphi(\mathbf{x})}{\partial\mathbf{s}} \|\mathbf{y} - \mathbf{x}\| \leqslant \varphi(\mathbf{y}) - \varphi(\mathbf{x}),$$

i.e.

$$\langle \varphi'(\mathbf{x}), \mathbf{s} \rangle \| \mathbf{y} - \mathbf{x} \| = \langle \varphi'(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \leqslant \varphi(\mathbf{y}) - \varphi(\mathbf{x}).$$

Now let the condition (2.15) be fulfilled. Consider the point $\mathbf{z} = \alpha\mathbf{x} + (1 - \alpha)\mathbf{y}$ for $0 \leqslant \alpha \leqslant 1$. Since $\mathbf{z} \in X$, we have

$$\langle \varphi'(\mathbf{z}), \mathbf{x} - \mathbf{z} \rangle \leqslant \varphi(\mathbf{x}) - \varphi(\mathbf{z}),$$

$$\langle \varphi'(\mathbf{z}), \mathbf{y} - \mathbf{z} \rangle \leqslant \varphi(\mathbf{y}) - \varphi(\mathbf{z}).$$

Multiplying the first and second inequalities by α and $(1 - \alpha)$, respectively, and adding together the results we find

$$0 = \langle \varphi'(\mathbf{z}), \mathbf{0} \rangle \leqslant \alpha\varphi(\mathbf{x}) + (1 - \alpha)\varphi(\mathbf{y}) - \varphi(\mathbf{z}). \quad \triangle$$

2.4.10. Extremal properties. We shall consider the problem of finding a point $\mathbf{x}^*$ in a convex set X, where a convex function $\varphi(\mathbf{x})$ on X attains minimum, i.e. the problem of finding an optimal point $\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$.

2.4.11. Theorem. *If $\varphi(\mathbf{x})$ is a convex function on a convex set X, then any point $\mathbf{x}^* \in X$ of local minimum is optimal for the problem of minimizing the function $\varphi(\mathbf{x})$ on the set X.*

Proof. Assume that $\mathbf{x}^*$ is not optimal, i.e. there is a point $\mathbf{x}' \in X$ such that

$$\varphi(\mathbf{x}') < \varphi(\mathbf{x}^*).$$

Consider the points

$$\mathbf{x} = \alpha\mathbf{x}' + (1 - \alpha)\mathbf{x}^*, \quad \alpha \in [0, 1].$$

Since the set X is convex, we have $\mathbf{x} \in X$. Further, the convexity of $\varphi(\mathbf{x})$ and the property of $\mathbf{x}'$ imply

$$\varphi(\mathbf{x}) = \varphi(\alpha\mathbf{x}' + (1 - \alpha)\mathbf{x}^*) \leqslant \alpha\varphi(\mathbf{x}')$$
$$+ (1 - \alpha)\varphi(\mathbf{x}^*) < \alpha\varphi(\mathbf{x}^*) + (1 - \alpha)\varphi(\mathbf{x}^*) = \varphi(\mathbf{x}^*),$$

i.e.

$$\varphi(\mathbf{x}) < \varphi(\mathbf{x}^*).$$

However, this contradicts the condition that $\mathbf{x}^*$ is a point of local minimum since for sufficiently small α the point $\mathbf{x}$ belongs to an arbitrarily small neighborhood of the point $\mathbf{x}^*$. $\triangle$

2.4.12. *If* $\varphi(\mathbf{x})$ *is a convex function on a convex set* X, *then the set of optimal points*

$$X^* = \{\mathbf{x}^* \in X: \ \varphi(\mathbf{x}^*) = \mu = \min_{\mathbf{x} \in X} \varphi(\mathbf{x})\}$$
$$= \operatorname{Arg\,min} \{\varphi(\mathbf{x}): \ \mathbf{x} \in X\}$$

is convex.

Proof. Let $\mathbf{x}', \mathbf{x}'' \in X^*$. Since $X^* \subseteq X$ and X is a convex set, we have

$$\mathbf{z} = \lambda \mathbf{x}' + (1 - \lambda)\, \mathbf{x}'' \in X$$

for any $\lambda \in [0, 1]$, and, in view of the convexity of $\varphi(\mathbf{x})$,

$$\varphi(\mathbf{z}) = \varphi(\lambda \mathbf{x}' + (1 - \lambda)\, \mathbf{x}'') \leqslant \lambda \varphi(\mathbf{x}') + (1 - \lambda)\, \varphi(\mathbf{x}'') = \lambda \mu + (1 - \lambda)\, \mu = \mu.$$

Moreover, $\varphi(\mathbf{z}) \geqslant \mu$, and therefore $\varphi(\mathbf{z}) = \mu$, i.e. $\mathbf{z} \in X^*$. $\triangle$

2.4.13. *If* $\varphi(\mathbf{x})$ *is strictly convex on a convex set* X *and* $\mathbf{x}^* \in X$ *is an optimal point, i.e.*

$$\mu = \varphi(\mathbf{x}^*) = \min_{\mathbf{x} \in X} \varphi(\mathbf{x}),$$

then

$$\varphi(\mathbf{x}) > \varphi(\mathbf{x}^*)$$

for all $\mathbf{x} \in X$ *and* $\mathbf{x} \neq \mathbf{x}^*$, *and, consequently, the point* $\mathbf{x}^*$ *is unique.*

Proof. Assume that there is a point $\mathbf{x}' \in X$, $\mathbf{x}' \neq \mathbf{x}^*$, such that

$$\varphi(\mathbf{x}') = \varphi(\mathbf{x}^*) = \mu.$$

Then, given any $\alpha \in (0, 1]$, the point $\mathbf{x} = \alpha \mathbf{x}' + (1 - \alpha)\, \mathbf{x}^*$ belongs to the set X, and, by the strong convexity of the function $\varphi(\mathbf{x})$, we have

$$\varphi(\mathbf{x}) = \varphi(\alpha \mathbf{x}' + (1 - \alpha)\, \mathbf{x}^*) < \alpha \varphi(\mathbf{x}') + (1 - \alpha)\, \varphi(\mathbf{x}^*) = \alpha \mu + (1 - \alpha)\, \mu = \mu,$$

which contradicts the optimality of the point $\mathbf{x}^*$. $\triangle$

2.4.14. A property of convex functions. *If* $\varphi(\mathbf{x})$ *is a convex function, then for any point* $\mathbf{z} \in [\mathbf{x}, \mathbf{y}]$ *there holds the inequality*

$$\frac{\varphi(\mathbf{y}) - \varphi(\mathbf{z})}{\|\mathbf{y} - \mathbf{z}\|} \geqslant \frac{\varphi(\mathbf{y}) - \varphi(\mathbf{x})}{\|\mathbf{y} - \mathbf{x}\|} \geqslant \frac{\varphi(\mathbf{z}) - \varphi(\mathbf{x})}{\|\mathbf{z} - \mathbf{x}\|}.$$

Proof. Assume that there is $\alpha \in (0, 1)$ such that for $z = \alpha x + (1 - \alpha) y$ holds

$$\frac{\varphi(y) - \varphi(z)}{\| y - z \|} < \frac{\varphi(y) - \varphi(x)}{\| y - x \|}.$$

But $\| y - z \| = \alpha \| y - x \|$, therefore

$$\frac{1}{\alpha}(\varphi(y) - \varphi(z)) < \varphi(y) - \varphi(x).$$

Thus we arrive at the inequality

$$\varphi(z) < \alpha\varphi(x) + (1 - \alpha)\varphi(y)$$

contradicting the convexity of the function $\varphi(x)$. $\triangle$

2.4.15. Theorem. *If X is a convex, closed, and unbounded set, then there is a direction s, $\| s \| = 1$, such that $x + \lambda s \in X$ for all $\lambda \geqslant 0$ and all $x \in X$.*

Proof. Since the set X is unbounded, there is a sequence $\{x_k\} \in X$, $\| x_k \| \to \infty$, $k \to +\infty$. Consider a fixed point $x \in X \setminus \{x_k\}$ and the direction $s_k = \| x_k - x \|^{-1}(x_k - x)$. It is obvious that there exists a subsequence $\{s_{k_i}\}$ such that $\lim\limits_{i \to \infty} s_{k_i} = s$. For any fixed $\lambda > 0$ there is an index k_0 such that $0 < \lambda \| x_{k_i} - x \|^{-1} < 1$ for all $k_i \geqslant k_0$.

Since

$$x + \lambda s_{k_i} = \frac{\lambda}{\| x_{k_i} - x \|} x_k + \left(1 - \frac{\lambda}{\| x_{k_i} - x \|}\right) x \in X$$

for all $k_i \geqslant k_0$, we see that, by virtue of the closedness of the set X, the point $\lim\limits_{i \to \infty} (x + \lambda s_{k_i}) = x + \lambda s$ belongs to X.

Further, let y be an arbitrary fixed element of X. Since $v_k = x + \lambda_k \lambda s \in X$ for all $\lambda \geqslant 1$, $\lambda_k > 1$, we have

$$\frac{1}{\lambda_k} v_k + \left(1 - \frac{1}{\lambda_k}\right) y = \frac{1}{\lambda_k}(x - y) + y + \lambda s \in X,$$

whence, for $\lambda_k \to \infty$, in view of the closedness of the set X, we obtain $y + \lambda s \in X$. $\triangle$

2.4.16. Theorem. *If $f(x)$ is a convex and continuous function on a closed convex set X, then for the set*

$$X(\beta) = \{x \in X: \ f(x) \leqslant \beta\}$$

to be bounded for any β *it is necessary and sufficient that there exist a number* α *such that the set* $X(\alpha) = \{x \in X: f(x) \leqslant \alpha\}$ *is nonempty and bounded.*

Proof. The necessity is obvious.

Sufficiency. Let the set $X(\alpha)$ be bounded for some α. If $\beta \leqslant \alpha$, then $X(\beta) \subset X(\alpha)$, and, consequently, $X(\beta)$ is a bounded set.

Assume that there is $\beta > \alpha$ such that the set $X(\beta)$ is unbounded. Then at a fixed point $x \in X(\alpha) \subset X(\beta)$ there is a direction s, $\|s\| = 1$, such that $x + \lambda s \in X(\beta)$ for all $\lambda \geqslant 0$. Since the set $X(\alpha)$ is bounded, there is $\lambda_0 > \operatorname{diam} X(\alpha)$ such that $x + \lambda s \notin X(\alpha)$ for all $\lambda \geqslant \lambda_0$. Thus, for $z = x + \lambda_0 s$, we have $f(x) \leqslant \alpha < f(z) \leqslant \beta$, and $f(v) \leqslant \beta$ for $v = v(\lambda) = x + \lambda\lambda_0 s$ and all $\lambda > 0$. Since $z = x + \lambda_0 s = \lambda^{-1}v + (1 - \lambda^{-1})x$, the convexity of the function $f(x)$ on the closed convex set $X(\beta)$ implies the inequality $f(z) \leqslant \lambda^{-1} f(v) + (1 - \lambda^{-1}) f(x)$ for $\lambda \geqslant 1$, whence $f(v) \geqslant \lambda (f(z) - f(x)) + f(x)$. However, $f(z) - f(x)$ is a fixed positive number, and therefore $f(v) \to +\infty$ for $\lambda \to +\infty$, which contradicts the condition that $f(v) \leqslant \beta$ for all $\lambda > 0$. $\triangle$

2.5. Strongly Convex Functions

2.5.1. We proved that if a convex function $\varphi(x)$ possesses a point of local minimum on a convex and closed set X, then this point is optimal, and, for a strictly convex function such a point is unique. We stress that these assertions are true only under the assumption that there exists a point of local minimum.

We now consider a class of functions for which on any nonempty closed set there always exists a point of minimum, and if, in addition, the set is convex, then this point is unique.

2.5.2. Definition. A function $\varphi(x)$ on a set X is said to be *strongly convex* if there is a constant $\rho > 0$ such that for any $x, y \in X$ such that $[x, y] \subset X$ and any $\alpha \in [0, 1]$ the inequality

$$\varphi(\alpha x + (1 - \alpha) y) \leqslant \alpha\varphi(x) + (1 - \alpha)\varphi(y) - \alpha(1 - \alpha)\rho \|x - y\|^2$$

is fulfilled.

In what follows the number ρ will be called the *parameter* of strong convexity.

2.5.3. An example of a strongly convex function. Consider a quadratic function

$$\varphi(\mathbf{x}) = \langle \mathbf{x}, B\mathbf{x} \rangle + \langle \mathbf{p}, \mathbf{x} \rangle,$$

where B is a strictly positive definite matrix. The strong convexity follows from the relation

$$\varphi(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}) = \alpha\varphi(\mathbf{x}) + (1-\alpha)\varphi(\mathbf{y}) - \alpha(1-\alpha)\langle \mathbf{x}-\mathbf{y}, B(\mathbf{x}-\mathbf{y}) \rangle$$

since

$$\langle \mathbf{x}-\mathbf{y}, B(\mathbf{x}-\mathbf{y}) \rangle \geqslant \lambda \| \mathbf{x}-\mathbf{y} \|^2,$$

where λ is the least eigenvalue of the matrix B.

We indicate some properties of strongly convex functions.

2.5.4. *If a function $\varphi(\mathbf{x})$ is strongly convex and continuous on a convex and closed set X, then for any point $\mathbf{y} \in X$ the set*

$$X_0 = \{\mathbf{x} \in X: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{y})\}$$

is bounded and there is a unique point

$$\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}.$$

Proof. We fix an arbitrary point $\mathbf{y} \in X$. By the continuity of the function $\varphi(\mathbf{x})$, for $\rho > 0$ there is $\varepsilon > 0$ such that $|\varphi(\mathbf{x}) - \varphi(\mathbf{y})| \leqslant \rho$ for all $\mathbf{x} \in U = \{\mathbf{x} \in X: \| \mathbf{x}-\mathbf{y} \| \leqslant \varepsilon\}$. Thus, $\varphi(\mathbf{x}) \geqslant \varphi(\mathbf{y}) - \rho$ for all $\mathbf{x} \in U$.

Let $\mathbf{x} \in X \setminus U$. Then $\alpha = \varepsilon \| \mathbf{x}-\mathbf{y} \|^{-1} < 1$. From 2.5.2 we derive $\alpha\varphi(\mathbf{x}) \geqslant \varphi(\mathbf{y} + \alpha(\mathbf{x}-\mathbf{y})) - (1-\alpha)\varphi(\mathbf{y}) + \alpha(1-\alpha)\rho \| \mathbf{x}-\mathbf{y} \|^2$. However, $\mathbf{z} = \mathbf{y} + \alpha(\mathbf{x}-\mathbf{y}) \in U$ since $\| \mathbf{z}-\mathbf{y} \| \leqslant \varepsilon$, and consequently, $\varphi(\mathbf{y} + \alpha(\mathbf{x}-\mathbf{y})) - \varphi(\mathbf{y}) \geqslant -\rho$. Therefore,

$$\alpha\varphi(\mathbf{x}) \geqslant -\rho + \alpha\varphi(\mathbf{y}) + \alpha(1-\alpha)\rho \| \mathbf{x}-\mathbf{y} \|^2,$$

whence

$$\varphi(\mathbf{x}) \geqslant \varphi(\mathbf{y}) + (1-\alpha)\rho \| \mathbf{x}-\mathbf{y} \|^2 - \frac{\rho}{\alpha} = \varphi(\mathbf{y}) - \rho\left(\varepsilon + \frac{1}{\varepsilon}\right) \| \mathbf{x}-\mathbf{y} \| + \rho \| \mathbf{x}-\mathbf{y} \|^2.$$

Hence, if the set X is unbounded, then $\varphi(\mathbf{x}) \to +\infty$ for $\| \mathbf{x} \| \to \infty$. It can now easily be seen that the set X_0

is bounded. Indeed, if otherwise, then there is a sequence $\{\mathbf{x}_k\} \subset X_0$ such that $\| \mathbf{x}_k \| \to \infty$ for $k \to \infty$. Therefore there exists an index $k_0 = k_0(\mathbf{y})$ beginning with which $\varphi(\mathbf{x}_k) > \varphi(\mathbf{y})$, $k \geqslant k_0$, and, consequently, $\mathbf{x}_k \notin X_0$ for $k \geqslant k_0$.

The existence of $\mathbf{x}^*$ is obvious since the continuity of $\varphi(\mathbf{x})$ on the bounded and closed set X_0 and the definition of X_0 imply that

$$\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X_0\} = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$$

As to the uniqueness of the point $\mathbf{x}^*$, it follows from the property that a strongly convex function is at the same time strictly convex. $\triangle$

2.5. *If $\varphi(\mathbf{x})$ is strongly convex on a convex and closed set X, then*

(a) *for all $\mathbf{x} \in X$ the inequality*

$$\| \mathbf{x} - \mathbf{x}^* \|^2 \leqslant \frac{2}{\rho} (\varphi(\mathbf{x}) - \varphi(\mathbf{x}^*))$$

is fulfilled, and if $\varphi(\mathbf{x}) \in C^1(X)$, then

(b) *for all $\mathbf{x}, \mathbf{y} \in X$ there hold*

$$\langle \varphi'(\mathbf{x}) - \varphi'(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geqslant \rho \| \mathbf{x} - \mathbf{y} \|^2,$$

(c) $\| \mathbf{x} - \mathbf{x}^* \| \leqslant \frac{1}{\rho} \| \varphi'(\mathbf{x}) \|$,

(d) $0 \leqslant \varphi(\mathbf{x}) - \varphi(\mathbf{x}^*) \leqslant \frac{1}{\rho} \| \varphi'(\mathbf{x}) \|^2$.

Proof. (a) The definition of strong convexity implies

$$\varphi\left(\frac{1}{2}\mathbf{x} + \frac{1}{2}\mathbf{x}^*\right) \leqslant \frac{1}{2}\varphi(\mathbf{x}) + \frac{1}{2}\varphi(\mathbf{x}^*) - \frac{1}{4}\rho \| \mathbf{x} - \mathbf{x}^* \|^2,$$

and, by virtue of

$$\varphi(\mathbf{x}^*) \leqslant \varphi\left(\frac{1}{2}\mathbf{x} + \frac{1}{2}\mathbf{x}^*\right),$$

the inequality in (a) holds.

(b) Since the function $\varphi(\mathbf{x})$ is convex, we have

$$\varphi(\mathbf{x}) - \varphi(\mathbf{y}) \leqslant \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle.$$

This and the definition 2.5.2 imply

$$\frac{1}{4}\rho \| x - y \|^2 \leqslant \frac{1}{2}\left[\varphi(x) - \varphi\left(\frac{1}{2}x + \frac{1}{2}y\right)\right]$$
$$+ \frac{1}{2}\left[\varphi(y) - \varphi\left(\frac{1}{2}x + \frac{1}{2}y\right)\right] \leqslant \frac{1}{4}\langle \varphi'(x),\ x - y\rangle$$
$$+ \frac{1}{4}\langle \varphi'(y),\ y - x\rangle = \frac{1}{4}\langle \varphi'(x) - \varphi'(y),\ x - y\rangle.$$

(c) At the points of minimum x^* of $\varphi(x)$ on X the inequality (see Sec. 3.4.7)

$$\langle \varphi'(x^*),\ x - x^*\rangle \geqslant 0$$

is fulfilled for all $x \in X$, and therefore it follows from (b) that

$$\rho \| x - x^* \|^2 \leqslant \langle \varphi'(x) - \varphi'(x^*),\ x - x^*\rangle$$
$$\leqslant \langle \varphi'(x),\ x - x^*\rangle \leqslant \| \varphi'(x) \| \| x - x^* \|,$$

i.e. the inequality in (c) is true.

(d) It follows from the convexity of $\varphi(x)$ and (c) that

$$0 \leqslant \varphi(x) - \varphi(x^*) \leqslant \langle \varphi'(x),\ x - x^*\rangle$$
$$\leqslant \| \varphi'(x) \| \| x - x^* \| \leqslant \frac{1}{\rho} \| \varphi'(x) \|^2. \quad \triangle$$

Chapter 3

FUNDAMENTALS OF MATHEMATICAL PROGRAMMING

In this chapter we shall consider conditions for the existence of local extrema of differentiable functions in feasible regions of a rather general type and also conditions for the existence of global extrema (minima) in convex programming problems.

3.1. Mathematical Programming Problems

3.1.1. The basic mathematical programming problem. Consider the set

$$X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, \quad i = \overline{1, m}\}, \tag{3.1}$$

where $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are given scalar functions. Let $\varphi(\mathbf{x})$ be a scalar function on the set X.

The problem of minimizing the function $\varphi(\mathbf{x})$ on the set X will be called the *basic mathematical programming problem.*

3.1.2. Notation. We shall agree that the expression

$$\varphi(\mathbf{x}) \to \min, \quad \mathbf{x} \in X, \tag{3.2}$$

or, equivalently,

$$\min\{\varphi(\mathbf{x}): \mathbf{x} \in X\} \text{ or } \min_{\mathbf{x} \in X} \varphi(\mathbf{x}),$$

symbolizes that one of the following problems is posed:

(1) find an optimal point $\mathbf{x}^* \in X$:

$$\varphi(\mathbf{x}^*) = \min_{\mathbf{x} \in X} \varphi(\mathbf{x}),$$

(2) if there is no such point $\mathbf{x}^*$, find

$$\varphi^* = \inf_{\mathbf{x} \in X} \varphi(\mathbf{x}),$$

(3) show that the function $\varphi(\mathbf{x})$ is unbounded from below on the set X,

(4) show that $X = \varnothing$.

3.1.3. If the set X is convex and the function $\varphi(\mathbf{x})$ is convex, (3.2) is called a *convex programming problem.*

3.1.4. The basic convex programming problem. It follows from 2.4.5 and 2.1.6 that for the set X (see (3.1)) to be convex it is sufficient that the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ be concave. If all the functions $f_i(\mathbf{x})$ in the convex programming problem (3.2) are concave and $\varphi(\mathbf{x})$ is convex, we shall call it the *basic convex programming problem.*

3.1.5. Terminology. The set X in the problem (3.2) is called the *feasible region*, the points belonging to it are called *feasible points*, and the inequalities $f_i(\mathbf{x}) \geqslant 0$ determining the feasible region are called *constraints.* The point $\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$ is called a *solution* or an *optimal point* or sometimes a *point of global minimum.* Finally, a point $\mathbf{x}$ where the necessary conditions for local minimum of the function $\varphi(\mathbf{x})$ on the set X are fulfilled will be called a *stationary point.*

3.2. Feasible Directions

3.2.1. Feasible directions. The notion of a feasible direction plays an important role in mathematical programming. In the present chapter feasible directions are an auxiliary means, however in what follows their part will be essential.

Definition. A direction $-\mathbf{s} \neq 0$ at a point $\mathbf{x} \in X$ is said to be *feasible* if there is a number $\bar{\beta} > 0$ such that $\mathbf{x} - \beta\mathbf{s} \in X$ for all $\beta \in [0, \bar{\beta}]$.

For example, if $X = \{\mathbf{x}: \mathbf{x} \geqslant 0\}$, then at the point $\mathbf{x} = 0$ any vector $-\mathbf{s} \geqslant 0$, $\mathbf{s} \neq 0$, determines a feasible direction, and at the point

$$\mathbf{x} = \begin{bmatrix} x_1 = 0 \\ x_2 > 0 \\ \cdots \\ x_n > 0 \end{bmatrix}$$

any direction $-\mathbf{s}$, where

$$\mathbf{s} = \begin{bmatrix} s_1 \leqslant 0 \\ s_2 \\ \cdots \\ s_n \end{bmatrix},$$

and $s_2, s_3, \ldots, s_n$ are arbitrary numbers, $\mathbf{s} \neq \mathbf{0}$, is a feasible direction.

It is obvious that if $\mathbf{x}$ is an interior point of a set X, then any vector $-\mathbf{s}$ determines a feasible direction at the point.

In subsequent sections we shall establish conditions for the existence of feasible directions.

3.2.2. Active constraints. It seems obvious that extremal properties of an objective function $\varphi(\mathbf{x})$ in a sufficiently small neighborhood of a fixed point $\mathbf{x} \in X$ are affected only by those constraints $f_i(\mathbf{x}) \geqslant 0$ for which the equations $f_i(\mathbf{x}) = 0$ are fulfilled at that point. In this connection, we introduce the following definition:

A constraint $f_i(\mathbf{x}) \geqslant 0$ is said to be *active* at a fixed point $\mathbf{x} \in X$ if $f_i(\mathbf{x}) = 0$.

3.2.3. In what follows we shall deal with a set of indices of active constraints at a point $\mathbf{x} \in X$:

$$I(\mathbf{x}) = \{i: \quad f_i(\mathbf{x}) = 0\}.$$

3.2.4. Assumptions. Throughout this chapter we shall assume that the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are continuously differentiable.

3.2.5. Let $\mathbf{s}$ be an n-dimensional vector and let σ be a scalar.

For a fixed $\mathbf{x} \in X$ we consider the following system of inequalities in the variables $\mathbf{s}$ and σ:

$$\langle f_i'(\mathbf{x}), \mathbf{s} \rangle + \sigma \leqslant 0, \quad i \in I(\mathbf{x}). \tag{3.3}$$

Under some conditions the solutions to this system determine the feasible directions at the point $\mathbf{x}$.

3.2.6. *If* $\mathbf{s}$ $(\|\mathbf{s}\| \neq 0)$ *satisfies the system* (3.3) *for some* $\sigma > 0$, *then* $-\mathbf{s}$ *is a feasible direction at the point* $\mathbf{x} \in X$.

Proof. It is natural to assume that $I(\mathbf{x}) \neq \varnothing$ since, if otherwise, $\mathbf{x}$ is an interior point of the set X, and therefore any direction $-\mathbf{s}$ is feasible. If $i \notin I(\mathbf{x})$, then $f_i(\mathbf{x}) > 0$

and a small displacement from the point $\mathbf{x}$ in any direction, in particular, in the direction $-\mathbf{s}$, does not violate this inequality. Let $i \in I(\mathbf{x})$. Assume that $f_i(\mathbf{x} - \beta\mathbf{s}) < 0$ for any $\beta > 0$, however small, i.e. $-\mathbf{s}$ is not a feasible direction. Since $f_i(\mathbf{x}) = 0$, we have

$$\frac{1}{\beta}[f_i(\mathbf{x}) - f_i(\mathbf{x} - \beta\mathbf{s})] > 0$$

for any $\beta > 0$, and, consequently (see (2.4.7)),

$$\lim_{\beta \to 0} \frac{1}{\beta}[f_i(\mathbf{x}) - f_i(\mathbf{x} - \beta\mathbf{s})] = \langle f_i'(\mathbf{x}), \mathbf{s}\rangle \geqslant 0,$$

which contradicts the condition (3.3) for $\sigma > 0$. $\triangle$

3.2.7. *If $-\mathbf{s}$ is a feasible direction at a point $\mathbf{x} \in X$, then there is $\sigma \geqslant 0$ such that the pair $\mathbf{s}, \sigma$ satisfies the system* (3.3).

Proof. Assume that there is at least one index $i \in I(\mathbf{x})$ such that $\langle f_i'(\mathbf{x}), \mathbf{s}\rangle > 0$. Since $-f_i(\mathbf{x} - \beta\mathbf{s}) = f_i(\mathbf{x}) - f_i(\mathbf{x} - \beta\mathbf{s}) = \beta\langle f_i'(\mathbf{x}), \mathbf{s}\rangle + o(\beta)$, we have $f_i(\mathbf{x} - \beta\mathbf{s}) < 0$ for sufficiently small $\beta > 0$, i.e. $\mathbf{x} - \beta\mathbf{s} \notin X$, which contradicts the assumption that $-\mathbf{s}$ is a feasible direction. $\triangle$

3.2.8. *If the set X is determined by a system of linear inequalities*:

$$X = \{\mathbf{x}: f_i(\mathbf{x}) = \langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \geqslant 0, \quad i = \overline{1, m}\},$$

then the conditions

$$\langle \mathbf{a}_i, \mathbf{s}\rangle \leqslant 0, \quad i \in I(\mathbf{x}),$$

are necessary and sufficient for $-\mathbf{s}$ to be a feasible direction at the point $\mathbf{x} \in X$.

Proof. We must find conditions under which the point $\mathbf{y} = \mathbf{x} - \beta\mathbf{s}$ belongs to the set X at least for sufficiently small $\beta > 0$:

$$\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i = \langle \mathbf{a}_i, \mathbf{x}\rangle - b_i - \beta\langle \mathbf{a}_i, \mathbf{s}\rangle.$$

If $i \notin I(\mathbf{x})$, then $\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i > 0$, and for sufficiently small β we have $\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i \geqslant 0$.

If $i \in I(\mathbf{x})$, then $\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i = -\beta\langle \mathbf{a}_i, \mathbf{s}\rangle$, and for the inequality $\langle \mathbf{a}_i, \mathbf{y}\rangle - b_i \geqslant 0$ to hold for $\beta > 0$ it is necessary and sufficient that $\langle \mathbf{a}_i, \mathbf{s}\rangle \leqslant 0$. $\triangle$

3.3. Extremal Properties

In the subsequent sections of this chapter we consider fundamentals of mathematical programming theory: existence theorems are proved for local extrema and solutions of mathematical programming problems.

In the present section we study existence conditions for stationary points of the basic mathematical programming problems, i.e., in accordance with the accepted terminology, sufficient conditions for the existence of local minima.

We once again stress that, as was indicated in 3.2.4, the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are assumed to be continuously differentiable.

3.3.1. In what follows the system of linear inequalities in the variables $\mathbf{s}$ and σ

$$\langle f_i'(\mathbf{x}), \mathbf{s}\rangle + \sigma \leqslant 0, \quad i \in I(\mathbf{x}), \tag{3.4}$$

$$-\langle \varphi'(\mathbf{x}), \mathbf{s}\rangle + \sigma \leqslant 0 \tag{3.5}$$

plays an important part.

3.3.2. Theorem. *For $\mathbf{x} \in X$ to be a point of local minimum of a function $\varphi(\mathbf{x})$ on the set X it is necessary that*

$$\sigma \leqslant 0 \tag{3.6}$$

in each pair $\mathbf{s}$, σ satisfying the system (3.4), (3.5).

Proof. Let $\mathbf{x}$ be a point of local minimum. Assume that there is a pair $\mathbf{s}$, σ satisfying the system (3.4), (3.5) such that $\sigma > 0$. According to (3.2.6), $-\mathbf{s}$ is a feasible direction at the point $\mathbf{x}$. By the continuity of $\varphi'(\mathbf{x})$ and the assumption that $\langle \varphi'(\mathbf{x}), \mathbf{s}\rangle \geqslant \sigma > 0$, for sufficiently small $\beta > 0$ we have $\langle \varphi'(\mathbf{x} - \beta\mathbf{s}), \mathbf{s}\rangle > 0$ and $\mathbf{x} - \beta\mathbf{s} \in X$. However, by the mean value theorem,

$$\varphi(\mathbf{x}) - \varphi(\mathbf{x} - \beta\mathbf{s}) = \beta \langle \varphi'(\mathbf{x} - \theta\beta\mathbf{s}), \mathbf{s}\rangle > 0, \quad 0 \leqslant \theta \leqslant 1.$$

Thus, in a local neighborhood of the point of local minimum $\mathbf{x}$ there is a point $\mathbf{y} = \mathbf{x} - \beta\mathbf{s}$ such that $\varphi(\mathbf{y}) < \varphi(\mathbf{x})$, which is a contradiction. $\triangle$

3.3.3. Following our terminology (see 3.1.5), Theorem 3.3.2 can be restated thus: if for each pair $\mathbf{s}$, σ satisfying the system (3.4), (3.5) the condition (3.6) is fulfilled,

then $\mathbf{x}$ is a stationary point in the basic mathematical programming problem.

3.3.4. Corollary. *If a point* $\mathbf{x} \in X$ *of local minimum of the function* $\varphi(\mathbf{x})$ *on the set* X *is such that* $I(\mathbf{x}) = \varnothing$, *then*

$$\varphi'(\mathbf{x}) = \mathbf{0}.$$

Proof. Assume that $\varphi'(\mathbf{x}) \neq \mathbf{0}$. We note that $\mathbf{x}$ is an interior point of the set X. Since $I(\mathbf{x}) = \varnothing$, only the single inequality (3.5) remains in the system (3.4), (3.5). The pair $\mathbf{s} = \varphi'(\mathbf{x})$, $\sigma = \langle \mathbf{s}, \mathbf{s} \rangle > 0$ satisfies the condition (3.5) whereas the condition (3.6) of Theorem 3.3.2 is violated, which is a contradiction. $\triangle$

3.3.5. Theorem. *If the system of vectors* $f_i'(\mathbf{x})$, $i \in I(\mathbf{x})$, *is linearly independent at the point* $\mathbf{x} \in X$ *of local minimum of the function* $\varphi(\mathbf{x})$ *on the set* X, *then there are numbers* $u_i \geqslant 0$, $i \in I(\mathbf{x})$, *such that*

$$\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x})} u_i f_i'(\mathbf{x}). \tag{3.7}$$

Proof. According to Theorem 3.3.2, the conditions (3.4), (3.5), and (3.6) hold. The condition (3.6) can be written formally as

$$\langle \mathbf{0}, \mathbf{s} \rangle + 1 \cdot \sigma \leqslant 0. \tag{3.8}$$

Applying the Farkas theorem to the system (3.4), (3.5), (3.8) we see that there are $v_i \geqslant 0$, $i \in I(\mathbf{x})$, and $v_0 \geqslant 0$ such that

$$\mathbf{0} = \sum_{i \in I(\mathbf{x})} v_i f_i'(\mathbf{x}) - v_0 \varphi'(\mathbf{x}), \tag{3.9}$$

$$1 = \sum_{i \in I(\mathbf{x})} v_i + v_0. \tag{3.10}$$

Since the assumption that $v_0 = 0$ contradicts the linear independence of the vectors $f_i'(\mathbf{x})$, $i \in I(\mathbf{x})$ (in view of (3.9) and (3.10)), we have $v_0 > 0$, and, consequently, for $u_i = v_0^{-1} v_i$, $i \in I(\mathbf{x})$, the condition (3.9) implies (3.7). $\triangle$

3.3.6. Remark. Theorem 3.3.5 is sometimes stated thus: if there are numbers $u_i \geqslant 0$ $(i = \overline{1, m})$ such that

$$\varphi'(\mathbf{x}) = \sum_{i=1}^{m} u_i f_i'(\mathbf{x}), \quad \sum_{i=1}^{m} u_i f_i(\mathbf{x}) = 0,$$

then $\mathbf{x}$ is a stationary point in the basic mathematical programming problem.

The validity of these conditions is obvious: it suffices to put $u_i = 0$ for $i \notin I(\mathbf{x})$.

3.3.7. Geometrical interpretation. The condition (3.7) can be given the following geometrical interpretation. Noting that the vectors $-f_i'(\mathbf{x})$, $i \in I(\mathbf{x})$, are outer normals to the boundary surfaces $f_i(\mathbf{x}) = 0$ at the point $\mathbf{x}$ and rewriting (3.7) in the form

$$-\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x})} u_i |-f_i'(\mathbf{x})|, \quad u_i \geqslant 0, \quad i \in I(\mathbf{x}),$$

we see that the condition (3.7) means that the antigradient can be represented as a linear combination, with nonnegative coefficients, of outer normals to the constraints at the point $\mathbf{x}$. In other words, the antigradient belongs to the cone spanned on the outer normals to the constraints at the point $\mathbf{x}$.

3.4. Extremal Properties on Convex Sets

3.4.1. In the case when the set $X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ is convex the conditions of linear independence of the vectors $f_i'(\mathbf{x})$ corresponding to the active constraints can be replaced in the foregoing theorem by a more easily verifiable condition, namely the so-called *regularity condition*. There are various regularity conditions for constraints; here we shall consider the following conditions.

3.4.2. Regularity condition. If for each $1 \leqslant i \leqslant m$ there is a point $\mathbf{x}_i \in X$ such that

$$f_i(\mathbf{x}) > 0, \tag{3.11}$$

then we say that the set X satisfies the regularity condition.

3.4.3. Slater's regularity condition: there exists a point $\mathbf{x} \in X$ such that

$$f_i(\mathbf{x}) > 0 \tag{3.12}$$

for all $i = \overline{1, m}$.

It can readily be proved that the conditions (3.11) and (3.12) are equivalent. It is obvious that (3.12) implies (3.11).

Now let (3.11) hold. Take

$$\mathbf{x}=\sum_{i=1}^{m}\alpha_i\mathbf{x}_i,\quad \sum_{i=1}^{m}\alpha_i=1,\quad \alpha_i\geqslant 0,\quad i=\overline{1,\ m}.$$

Then (3.12) follows from Jensen's inequality 2.4.6 for concave functions $f_i(\mathbf{x})$.

3.4.4. Theorem. *If $f_i(\mathbf{x})$ are concave functions, the set $X=\{\mathbf{x}:\ f_i(\mathbf{x})\geqslant 0,\ i=\overline{1,\ m}\}$ is regular in Slater's sense, and $\mathbf{x}\in X$ is a point of local minimum of a function $\varphi(\mathbf{x})$ on X, then there are numbers $u_i\geqslant 0$, $i\in I(\mathbf{x})$, such that*

$$\varphi'(\mathbf{x})=\sum_{i\in I(\mathbf{x})}u_i f_i'(\mathbf{x}).$$

Proof. Repeating the proof of Theorem 3.3.5 we arrive at

$$v_i\geqslant 0,\ i\in I(\mathbf{x}),\ v_0\geqslant 0,$$

$$\mathbf{0}=\sum_{i\in I(\mathbf{x})}v_i f_i'(\mathbf{x})-v_0\varphi'(\mathbf{x}), \tag{3.13}$$

$$1=\sum_{i\in I(\mathbf{x})}v_i+v_0. \tag{3.14}$$

To complete the proof it remains to show that $v_0>0$. Assume that $v_0=0$. Then (3.14) implies that there is at least one $v_l>0$, $l\in I(\mathbf{x})$. By the regularity of the set X, it follows that there is a point $\mathbf{z}\in X$ such that $f_i(\mathbf{z})>0$ ($i=\overline{1,\ m}$). Therefore, $-\mathbf{s}=\mathbf{z}-\mathbf{x}$ is a feasible direction. Since $f_l(\mathbf{x})$ is a concave function, 2.4.9 implies

$$-\langle f_l'(\mathbf{x}),\ \mathbf{s}\rangle\geqslant f_l(\mathbf{z})-f_l(\mathbf{x})>0.$$

Multiplying scalarly (3.13) by $-\mathbf{s}$ we find:

$$0=\sum_{i\in I(\mathbf{x})}v_i\langle f_i'(\mathbf{x}),\ \mathbf{s}\rangle. \tag{3.15}$$

Since $-\mathbf{s}$ is a feasible direction at the point $\mathbf{x}$, Theorem 3.2.7 implies that $\langle f_i'(\mathbf{x}),\ \mathbf{s}\rangle\leqslant 0$, $i\in I(\mathbf{x})$. Thus, all the terms on the right-hand side of (3.15) are of the same sign, and one of them, namely the lth, is sure to be non-zero: $-v_l\langle f_l'(\mathbf{x}),\ \mathbf{s}\rangle>0$, which contradicts the fact that the whole sum is equal to zero. $\triangle$

Remark. In the case of convex functions $f_i(\mathbf{x})$ ($i=\overline{1,\ m}$) the regularity condition becomes unnecessary.

Indeed, *if* $f_i(\mathbf{x})$ *are convex functions, then*

$$\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x})} u_i f_i'(\mathbf{x}), \quad u_i \geqslant 0,$$

is a necessary condition for $\mathbf{x}$ *to be a point of local minimum of the function* $\varphi(\mathbf{x})$ *on the set* X.

This assertion follows from the Farkas theorem if we show that for any $\mathbf{u} \in E_n$ such that

$$\langle f_i'(\mathbf{x}), \mathbf{u} \rangle \geqslant 0, \quad i \in I(\mathbf{x}),$$

the inequality

$$\langle \varphi'(\mathbf{x}), \mathbf{u} \rangle \geqslant 0$$

holds.

Since the length of the vector $\mathbf{u}$ does not affect the validity of this inequality, it suffices to prove the inequality for vectors of a sufficiently small length. Let $\mathbf{u} = \mathbf{z} - \mathbf{x}$ and $\mathbf{z}$ belong to a sufficiently small neighborhood of $\mathbf{x}$ such that $f_j(\mathbf{z}) > 0$, $j \in \{1, \ldots, n\} \setminus I(\mathbf{x})$. The point $\mathbf{z}$ belongs to X since

$$f_i(\mathbf{z}) = f_i(\mathbf{z}) - f_i(\mathbf{x}) \geqslant \langle f_i'(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle \geqslant 0, \quad i \in I(\mathbf{x}).$$

If $\langle \varphi'(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle < 0$, then, by Taylor's formula,

$$\varphi(\mathbf{z}) - \varphi(\mathbf{x}) = \langle \varphi'(\mathbf{x}), \mathbf{z} - \mathbf{x} \rangle + o(\| \mathbf{z} - \mathbf{x} \|) < 0$$

since, by the hypothesis, $\| \mathbf{z} - \mathbf{x} \|$ is sufficiently small, which contradicts the condition that $\mathbf{x}$ is a point of local minimum.

3.4.5. Theorem. *If* $f_i(\mathbf{x})$ *are concave functions, the closed set* $X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ *is regular in Slater's sense, and* $\mathbf{x} \in X$ *is a point of local minimum of a function* $\varphi(\mathbf{x})$ *on the set* X, *then*

$$\mathbf{x} = \mathbf{p}(\mathbf{x} - \varphi'(\mathbf{x})).$$

Here $\mathbf{p}(\mathbf{v})$ denotes the projection of the point $\mathbf{v}$ on the set X.

Proof. Let $\mathbf{y}$ be an arbitrary point belonging to the set X. Then $-\mathbf{s} = \mathbf{y} - \mathbf{x}$ is a feasible direction at the point $\mathbf{x}$. We conclude from Theorem 3.4.4 that

$$\langle (\mathbf{x} - \varphi'(\mathbf{x})) - \mathbf{x}, \mathbf{y} - \mathbf{x} \rangle = \langle \varphi'(\mathbf{x}), \mathbf{s} \rangle = \sum_{i \in I(\mathbf{x})} u_i \langle f_i'(\mathbf{x}), \mathbf{s} \rangle \leqslant 0.$$

By Theorem 2.2.3, it follows that $\mathbf{x}$ is the projection of the point $\mathbf{x} - \varphi'(\mathbf{x})$ on the set X. $\triangle$

3.4.6. The case of linear constraints. If the constraints determining the feasible region are linear, then the foregoing theorem is valid without the assumption that the set X is regular.

Theorem. *If* $f_i(\mathbf{x})$ $(i = \overline{1, m})$ *are linear functions and*

$$\mathbf{x} \in X = \{\mathbf{x}: f_i(\mathbf{x}) = \langle \mathbf{a}_i, \mathbf{x} \rangle - b_i \geqslant 0, \quad i = \overline{1, m}\}$$

is a point of local minimum of a function $\varphi(\mathbf{x})$ *on a set* X, *then there are numbers* $u_i \geqslant 0$, $i \in I(\mathbf{x})$, *such that*

$$\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x})} u_i \mathbf{a}_i.$$

Proof. Let $\varepsilon > 0$ be so small that for all the points belonging to the neighborhood

$$U_\varepsilon(\mathbf{x}) = \{\mathbf{y} \in X: \| \mathbf{y} - \mathbf{x} \| \leqslant \varepsilon\}$$

of the point $\mathbf{x}$ the inequality

$$\varphi(\mathbf{y}) \geqslant \varphi(\mathbf{x})$$

is fulfilled. Consider an arbitrary point $\mathbf{z} \neq \mathbf{x}$ belonging to the set X. In view of the convexity of the set X, we have $\mathbf{x} - \beta(\mathbf{x} - \mathbf{z}) \in U_\varepsilon(\mathbf{x})$ for all $\beta \in (0, \bar{\beta}]$, where

$$\bar{\beta} = \min\left\{1, \frac{\varepsilon}{\| \mathbf{z} - \mathbf{x} \|}\right\}.$$

Therefore,

$$\lim_{\beta \to +0} \frac{\varphi(\mathbf{x}) - \varphi(\mathbf{x} - \beta(\mathbf{x} - \mathbf{z}))}{\beta} = \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{z} \rangle \leqslant 0.$$

If we put $\mathbf{s} = \mathbf{x} - \mathbf{z}$, then $-\mathbf{s}$ is a feasible direction at the point $\mathbf{x}$, and the above inequality takes the form

$$\langle \varphi'(\mathbf{x}), \mathbf{s} \rangle \leqslant 0. \tag{3.16}$$

Since $-\mathbf{s}$ is an arbitrary feasible direction at the point $\mathbf{x}$, it follows from Theorem 3.2.8 that the inequality (3.16) must hold for all $\mathbf{s}$ satisfying

$$\langle \mathbf{a}_i, \mathbf{s} \rangle \leqslant 0, \quad i \in I(\mathbf{x}). \tag{3.17}$$

Applying Theorem 2.3.3 to (3.16) and (3.17) we immediately prove the theorem. $\triangle$

3.4.7. When deriving the condition (3.16) we used only the property of convexity of the set X, and therefore we

can state the following condition for stationarity of a point $\mathbf{x}$ in the convex programming problem: *for a point* $\mathbf{x}$ *belonging to a convex set* X *to be a point of local minimum of a function* $\varphi(\mathbf{x})$ *on* X *it is necessary that the derivatives in all feasible directions at that point be nonpositive*:

$$\frac{\partial \varphi(\mathbf{x})}{\partial (-\mathbf{s})} \leqslant 0.$$

3.5. Sifficient Conditions for Optimality

3.5.1. Theorem. *For* $\mathbf{x} \in X$ *to be a point of global minimum in the basic convex programming problem* 3.1.4 *it is sufficient that there exist numbers* $u_i \geqslant 0$, $i \in I(\mathbf{x})$, *such that*

$$\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x})} u_i f_i'(\mathbf{x}).$$

Proof. As a consequence of the convexity of the set X, $-\mathbf{s} = \mathbf{y} - \mathbf{x}$ is a feasible direction at the point $\mathbf{x}$ for any $\mathbf{y} \in X$. It follows from Theorem 3.2.7 that $\langle f_i'(\mathbf{x}), \mathbf{s} \rangle \leqslant 0$, $i \in I(\mathbf{x})$, and, since $\varphi(\mathbf{x})$ is a convex function, using (2.15) we derive the relation

$$\varphi(\mathbf{x}) - \varphi(\mathbf{y}) \leqslant \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle = \langle \varphi'(\mathbf{x}), \mathbf{s} \rangle$$
$$= \left\langle \sum_{i \in I(\mathbf{x})} u_i f_i'(\mathbf{x}), \mathbf{s} \right\rangle = \sum_{i \in I(\mathbf{x})} u_i \langle f_i'(\mathbf{x}), \mathbf{s} \rangle \leqslant 0$$

holding for any $\mathbf{y} \in X$. $\triangle$

3.5.2. The Kuhn-Tucker theorem (differential form). Theorems 3.4.4 and 3.5.1 can be combined to state the optimality criterion: *if* $f_i(\mathbf{x})$ *are concave functions*, $\varphi(\mathbf{x})$ *is a convex function, and the set* $X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ *is regular in Slater's sense, then for* $\mathbf{x} \in X$ *to be an optimal point it is necessary and sufficient that there exist numbers* $u_i \geqslant 0$ $(i = \overline{1, m})$ *such that*

$$\varphi'(\mathbf{x}) = \sum_{i=1}^{m} u_i f_i'(\mathbf{x}), \quad \sum_{i=1}^{m} u_i f_i(\mathbf{x}) = 0.$$

To verify the validity of the theorem it suffices to recall that a point of local minimum of a convex function is optimal (Theorem 2.4.11). $\triangle$

3.5.3. Among the constraints determining the feasible region X those requiring the nonnegativity of feasible points are often separated out:

$$X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, \quad i = \overline{1, m}, \quad \mathbf{x} \geqslant \mathbf{0}\}.$$

In this case the Kuhn-Tucker theorem is formulated thus: *if the functions* $f_i(\mathbf{x})$ *are concave, the function* $\varphi(\mathbf{x})$ *is convex, and the set* X *is regular in Slater's sense, then for* $\mathbf{x} \in X$ *to be an optimal point it is necessary and sufficient that there exist numbers* $u_i \geqslant 0$ $(i = \overline{1, m})$ *and* $v_j \geqslant 0$ $(j = \overline{1, n})$ *such that*

$$\varphi'(\mathbf{x}) = \sum_{i=1}^{m} u_i f'(\mathbf{x}) + \sum_{j=1}^{n} v_j \mathbf{e}_j,$$

$$\sum_{i=1}^{m} u_i f_i(\mathbf{x}) = 0, \quad \sum_{j=1}^{n} v_j x_j = 0.$$

Here $\mathbf{e}_j$ $(j = \overline{1, n})$ is the jth coordinate vector: $\mathbf{e}_j^T = (0, \ldots, 1, \ldots, 0)$, where unity occupies the jth place.

3.5.4. The case of linear constraints (optimality criterion). Theorems 3.4.6 and 3.5.1 can also be combined to state *the optimality criterion: for* $\mathbf{x}$ *to be a point of global minimum of a convex function* $\varphi(\mathbf{x})$ *on the set* $X = \{\mathbf{x}: \langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \geqslant 0, i = \overline{1, m}\}$ *it is necessary and sufficient that there exist numbers* $u_i \geqslant 0$ $(i = \overline{1, m})$ *such that*

$$\varphi'(\mathbf{x}) = \sum_{i=1}^{m} u_i \mathbf{a}_i, \quad \sum_{i=1}^{m} u_i (\langle \mathbf{a}_i, \mathbf{x}\rangle - b_i) = 0.$$

3.5.5. Theorem. *For* $\mathbf{x} \in X$ *to be a point of global minimum in the basic convex programming problem* 3.1.4 *it is sufficient that for all* $\mathbf{s}$ *satisfying the system of inequalities*

$$\langle f_i'(\mathbf{x}), \mathbf{s}\rangle \leqslant 0, \quad i \in I(\mathbf{x}) \tag{3.18}$$

the condition

$$\langle \varphi'(\mathbf{x}), \mathbf{s}\rangle \leqslant 0 \tag{3.19}$$

hold.

Proof. Applying the Farkas theorem 2.3.3 to (3.18) and (3.19) we immediately arrive at the conditions of Theorem 3.5.1. △

3.6. Lagrangian Function. Optimality Conditions

3.6.1. Saddle point. Consider an n-dimensional vector $\mathbf{x}$ belonging to a convex set Γ and a nonnegative m-dimensional vector $\mathbf{y} \geqslant \mathbf{0}$. Let a function $L(\mathbf{x}, \mathbf{y})$ be convex with respect to $\mathbf{x}$ on the set Γ and concave with respect to $\mathbf{y}$ on the nonnegative orthant.

Definition. A pair $\mathbf{x}^*$, $\mathbf{y}^*$ is called a *saddle point* of the function $L(\mathbf{x}, \mathbf{y})$ on the set of all $\mathbf{x} \in \Gamma$ and $\mathbf{y} \geqslant \mathbf{0}$ if

$$\mathbf{x}^* \in \Gamma, \quad \mathbf{y}^* \geqslant \mathbf{0},$$
$$L(\mathbf{x}^*, \mathbf{y}) \leqslant L(\mathbf{x}^*, \mathbf{y}^*) \leqslant L(\mathbf{x}, \mathbf{y}^*) \qquad (3.20)$$

for all $\mathbf{x} \in \Gamma$ and $\mathbf{y} \geqslant \mathbf{0}$.

The inequalities (3.20) can also be written in the following way:

$$L(\mathbf{x}^*, \mathbf{y}^*) = \min_{\mathbf{x} \in \Gamma} \max_{\mathbf{y} \geqslant 0} L(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{y} \geqslant 0} \min_{\mathbf{x} \in \Gamma} L(\mathbf{x}, \mathbf{y}).$$

3.6.2. Conditions for the existence of a saddle point. For our further aims it suffices to consider the cases where

$$\Gamma = E_n^+ = \{\mathbf{x}: \mathbf{x} \geqslant \mathbf{0}\} \text{ and } \Gamma = E_n.$$

Let a function $L(\mathbf{x}, \mathbf{y})$ be convex with respect to $\mathbf{x}$ for all $\mathbf{x} \geqslant \mathbf{0}$, concave with respect to $\mathbf{y}$ for all $\mathbf{y} \geqslant \mathbf{0}$, and continuously differentiable with respect to $\mathbf{x}$ and $\mathbf{y}$.

Theorem. *For a pair* $\mathbf{x}^*$, $\mathbf{y}^*$ $(\mathbf{x}^* \geqslant \mathbf{0}, \mathbf{y}^* \geqslant \mathbf{0})$ *to be a saddle point of the function* $L(\mathbf{x}, \mathbf{y})$ *in the region* $\mathbf{x} \geqslant \mathbf{0}$, $\mathbf{y} \geqslant \mathbf{0}$ *it is necessary and sufficient that the conditions*

$$\frac{\partial L^*}{\partial \mathbf{x}} \geqslant \mathbf{0}, \qquad (3.21)$$

$$\left\langle \mathbf{x}^*, \frac{\partial L^*}{\partial \mathbf{x}} \right\rangle = 0, \qquad (3.22)$$

$$\mathbf{x}^* \geqslant \mathbf{0} \qquad (3.23)$$

$$\frac{\partial L^*}{\partial \mathbf{y}} \leqslant \mathbf{0}, \qquad (3.24)$$

$$\left\langle \mathbf{y}^*, \frac{\partial L^*}{\partial \mathbf{y}} \right\rangle = 0, \qquad (3.25)$$

$$\mathbf{y}^* \geqslant \mathbf{0} \qquad (3.26)$$

be fulfilled, where

$$\frac{\partial L^*}{\partial \mathbf{x}} \triangleq \frac{\partial L(\mathbf{x}, \mathbf{y})}{\partial \mathbf{x}}\Big|_{\substack{\mathbf{x}=\mathbf{x}^* \\ \mathbf{y}=\mathbf{y}^*}}, \quad \frac{\partial L^*}{\partial \mathbf{y}} \triangleq \frac{\partial L(\mathbf{x}, \mathbf{y})}{\partial \mathbf{y}}\Big|_{\substack{\mathbf{x}=\mathbf{x}^* \\ \mathbf{y}=\mathbf{y}^*}}.$$

Proof. We write (3.21)-(3.26) in the following equivalent form:

$$\frac{\partial L^*}{\partial x_i} \geqslant 0, \quad i = \overline{1, n}, \tag{3.21'}$$

$$x_i^* \frac{\partial L^*}{\partial x_i} = 0, \quad i = \overline{1, n}, \tag{3.22'}$$

$$x_i^* \geqslant 0, \quad i = \overline{1, n}, \tag{3.23'}$$

$$\frac{\partial L^*}{\partial y_j} \leqslant 0, \quad j = \overline{1, m}, \tag{3.24'}$$

$$y_j^* \frac{\partial L^*}{\partial y_j} = 0, \quad j = \overline{1, m}, \tag{3.25'}$$

$$y_j^* \geqslant 0, \quad j = \overline{1, m}. \tag{3.26'}$$

Necessity. Let the conditions (3.20) hold for $\Gamma = \{\mathbf{x}: \mathbf{x} \geqslant \mathbf{0}\}$. In particular, it follows that

$$L(x_i, \mathbf{y}^*) \triangleq L(x_1^*, \ldots, x_{i-1}^*, x_i, x_{i+1}^*, \ldots, x_n^*, \mathbf{y}^*) \geqslant L(\mathbf{x}^*, \mathbf{y}^*)$$

for all $x_i \geqslant 0$, i.e. x_i^* is a point of minimum of the convex function of one variable $L(x_i, \mathbf{y}_i^*)$ on the half-line $x_i \geqslant 0$. The relations (3.21')-(3.23') are necessary conditions for minimum, in particular, for local minimum, for the function of one variable for $x_i \geqslant 0$ $\left(\text{since either } x_i^* \text{ is an interior point of the semiaxis } x_i \geqslant 0 \text{ and then } \frac{\partial L^*}{\partial x_i} = 0, \text{ or } x_i^* = 0, \text{ and then } \frac{\partial L^*}{\partial x_i} \geqslant 0\right)$. Similarly, based on the concavity of the function $L(\mathbf{x}, \mathbf{y})$ with respect to $\mathbf{y}$, we can easily prove the validity of the conditions (3.24')-(3.26').

Sufficiency. Let the conditions (3.21)-(3.26) be fulfilled. Since $L(\mathbf{x}, \mathbf{y})$ is convex with respect to $\mathbf{x}$ for $\mathbf{x} \geqslant \mathbf{0}$, using the inequality in (2.4.9) we obtain

$$L(\mathbf{x}, \mathbf{y}^*) \geqslant L(\mathbf{x}^*, \mathbf{y}^*) + \left\langle \mathbf{x} - \mathbf{x}^*, \frac{\partial L^*}{\partial \mathbf{x}} \right\rangle$$

which, together with (3.21)-(3.23), implies

$$L(\mathbf{x}^*, \mathbf{y}^*) \leqslant L(\mathbf{x}, \mathbf{y}^*), \quad \mathbf{x} \geqslant \mathbf{0}.$$

The left-hand inequality in (3.20) is proved in a similar way. △

3.6.3. If $\Gamma = E_n$, an analogous argument shows that the saddle point is determined by

$$\frac{\partial L^*}{\partial x} = \mathbf{0}$$

and the conditions (3.24)-(3.26).

3.6.4. The Lagrangian function. Let $\mathbf{f}(\mathbf{x})$ be an m-dimensional vector: $\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$. Consider the following convex programming problem:

$$\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}, \quad X = \{\mathbf{x} \in \Gamma: \mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}\}. \quad (3.27)$$

Here Γ is a convex and closed set, $\varphi(\mathbf{x})$ is a convex function, and all the functions $f_i(\mathbf{x})$ are concave. We note that, according to the terminology, for $\Gamma = E_n$ the problem (3.27) is the basic convex programming problem.

Definition. The function

$$L(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle \quad (3.28)$$

defined for all $\mathbf{x} \in E_n$ and $\mathbf{y} \geqslant \mathbf{0}$ is called the *Lagrangian function* for the convex programming problem (3.27).

3.6.5. In classical problems on conditional extremum (where the feasible region is determined by a system of equations) an important role is played by the Lagrange multiplier method: the solution to the original problem is sought among the stationary points of the function $L(\mathbf{x}, \mathbf{y})$, i.e. points satisfying the system of equations

$$\frac{\partial L}{\partial \mathbf{x}} = \mathbf{0}, \quad \frac{\partial L}{\partial \mathbf{y}} = \mathbf{0}.$$

In convex (and, in particular, linear) programming problems the Lagrangian function is also very important: under some rather general assumptions the convex programming problem is reduced to the determination of saddle points of the Lagrangian function.

3.6.6. Sufficient conditions for optimality. Theorem. *If a pair* $\mathbf{x}^*, \mathbf{y}^*$ *is a saddle point of the Lagrangian function* (3.28) *on the set* $\mathbf{x} \in \Gamma$, $\mathbf{y} \geqslant \mathbf{0}$, *then* $\mathbf{x}^*$ *is an optimal point in the convex programming problem* (3.27).

Proof. (3.28) and the definition of a saddle point (3.20) imply the inequalities

$$\varphi(\mathbf{x}^*) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}^*)\rangle \leqslant \varphi(\mathbf{x}^*) - \langle \mathbf{y}^*, \mathbf{f}(\mathbf{x}^*)\rangle \leqslant \varphi(\mathbf{x}) - \langle \mathbf{y}^*, \mathbf{f}(\mathbf{x})\rangle \tag{3.29}$$

holding for all $\mathbf{x} \in \Gamma$, $\mathbf{y} \geqslant \mathbf{0}$. It follows from the left-hand inequality that

$$\langle \mathbf{y}, \mathbf{f}(\mathbf{x}^*)\rangle \geqslant \langle \mathbf{y}^*, \mathbf{f}(\mathbf{x}^*)\rangle, \tag{3.30}$$

and, since $\mathbf{y}^* \geqslant \mathbf{0}$ and the inequality (3.30) is valid for any $\mathbf{y} \geqslant \mathbf{0}$, we have $\mathbf{f}(\mathbf{x}^*) \geqslant \mathbf{0}$.

In particular, (3.30) holds for $\mathbf{y} = \mathbf{0}$ as well, and therefore

$$\langle \mathbf{y}^*, \mathbf{f}(\mathbf{x}^*)\rangle \leqslant 0,$$

and, consequently (since $\mathbf{y}^* \geqslant \mathbf{0}$ and $\mathbf{f}(\mathbf{x}^*) \geqslant \mathbf{0}$),

$$\langle \mathbf{y}^*, \mathbf{f}(\mathbf{x}^*)\rangle = 0. \tag{3.31}$$

If $\mathbf{x} \in X$, then (3.27) implies that $\mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}$, and therefore for $\mathbf{x} \in X$ we have

$$\langle \mathbf{y}^*, \mathbf{f}(\mathbf{x})\rangle \geqslant 0. \tag{3.32}$$

Since the inequality (3.29) is fulfilled for all $\mathbf{x} \in \Gamma$ and, in particular, for $\mathbf{x} \in X$, the right-hand inequality (3.29) and also (3.31) and (3.32) imply that

$$\varphi(\mathbf{x}^*) \leqslant \varphi(\mathbf{x}) - \langle \mathbf{y}^*, \mathbf{f}(\mathbf{x})\rangle \leqslant \varphi(\mathbf{x})$$

for all $\mathbf{x} \in X$. However, $\mathbf{x}^* \in X$ (since $\mathbf{x}^* \in \Gamma$ and $\mathbf{f}(\mathbf{x}^*) \geqslant \mathbf{0}$), and, consequently, $\mathbf{x}^*$ is an optimal point. $\triangle$

Remark. In the proof of the theorem we used neither the properties of convexity of the function $\varphi(\mathbf{x})$ and the set Γ nor the concavity of the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ nor any smoothness conditions. Therefore the existence of a saddle point $\mathbf{x}^*$, $\mathbf{y}^*$ of the Lagrangian function determines the optimality of the point $\mathbf{x}^*$ for the general mathematical programming program. It should be stressed that the converse assertion that the optimality of the point $\mathbf{x}^*$ implies the existence of a saddle point $\mathbf{x}^*$, $\mathbf{y}^*$ of the Lagrangian function is true only for the convex programming problem and, besides, under the condition of regularity of the feasible region. This is the well-known

Kuhn-Tucker theorem. It will be proved below, under the assumption that the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are continuously differentiable, as an obvious consequence of Theorems 3.5.2 and 3.6.2.

3.6.7. The Kuhn-Tucker theorem. Let the feasible region in the convex programming problem have the form

$$X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0,\ i = \overline{1, m},\ \mathbf{x} \geqslant 0\}.$$

We shall assume that the convex function $\varphi(\mathbf{x})$ and the concave functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are continuously differentiable.

If the set X in the convex programming problem

$$\min\{\varphi(\mathbf{x}): \mathbf{x} \in X\},\quad X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0,\ i = \overline{1, m},\ \mathbf{x} \geqslant 0\}$$

possesses the property of regularity 3.4.2, *then the existence of* $\mathbf{y}^* \geqslant 0$ *such that the pair* $\mathbf{x}^*$, $\mathbf{y}^*$ *is a saddle point of the Lagrangian function*

$$L(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle$$

on the set $\mathbf{x} \geqslant 0$, $\mathbf{y} \geqslant 0$ *is a necessary and sufficient condition for optimality of the point* $\mathbf{x}^* \in X$.

Proof. The sufficiency follows from Theorem 3.6.6 for $\Gamma = \{\mathbf{x}: \mathbf{x} \geqslant 0\}$.

Necessity. Let us prove that the conditions of Theorem 3.5.3:

$$\varphi'(\mathbf{x}^*) = \sum_{i=1}^{m} y_i^* f_i'(\mathbf{x}^*) + \sum_{j=1}^{n} v_j^* \mathbf{e}_j, \tag{3.33}$$

$$\sum_{i=1}^{m} y_i^* f_i(\mathbf{x}^*) = 0, \tag{3.34}$$

$$y_i^* \geqslant 0,\ i = \overline{1, m}, \tag{3.35}$$

$$\sum_{j=1}^{n} v_j^* x_j^* = 0, \tag{3.36}$$

$$v_j^* \geqslant 0,\ j = \overline{1, n}, \tag{3.37}$$

and the conditions of Theorem 3.6.2 determining a saddle point of the function $L(\mathbf{x}, \mathbf{y})$:

$$\frac{\partial L^*}{\partial \mathbf{x}} \geqslant \mathbf{0}, \tag{3.38}$$

$$\left\langle \mathbf{x}^*, \frac{\partial L^*}{\partial \mathbf{x}} \right\rangle = 0, \tag{3.39}$$

$$\mathbf{x}^* \geqslant \mathbf{0}, \tag{3.40}$$

$$\frac{\partial L^*}{\partial y} \leqslant \mathbf{0}, \tag{3.41}$$

$$\left\langle \mathbf{y}^*, \frac{\partial L^*}{\partial \mathbf{y}} \right\rangle = 0, \tag{3.42}$$

$$\mathbf{y}^* \geqslant \mathbf{0}, \tag{3.43}$$

are equivalent. The equivalence will be expressed by the symbol $\sim$.

Let $\mathbf{v}^T = (v_1^*, v_2^*, \ldots, v_n^*)$ and $\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, \mathbf{f}_m(\mathbf{x}))$. We note that

$$\frac{\partial L^*}{\partial \mathbf{x}} = \varphi'(\mathbf{x}^*) - \sum_{i=1}^{m} y_i^* \mathbf{f}_i'(\mathbf{x}^*) \underline{\underline{\triangle}} \mathbf{v},$$

$$\frac{\partial L^*}{\partial \mathbf{y}} = -\mathbf{f}(\mathbf{x}^*).$$

It is obvious that

$$(3.38) \sim (3.37),$$
$$(3.39) \sim (3.36),$$
$$(3.42) \sim (3.34),$$
$$(3.43) \sim (3.35).$$

The conditions (3.40) and (3.41) imply that $\mathbf{x}^* \in X$. $\triangle$

3.6.8. For the basic convex programming problem 3.1.4 the Kuhn-Tucker theorem is stated in the following way: *if the set* $X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ *in the basic convex programming problem possesses the regularity property* 3.4.2, *then the existence of* $\mathbf{y}^* \geqslant \mathbf{0}$ *such that the pair* $\mathbf{x}^*$, $\mathbf{y}^*$ *is a saddle point of the Lagrangian function on the set* $\mathbf{x} \in E_n$, $\mathbf{y} \geqslant \mathbf{0}$ *is a necessary and sufficient condition for optimality of the point* $\mathbf{x}^* \in X$.

The theorem is implied by 3.5.2 and 3.6.3.

3.6.9. The case of linear constraints. *If the function* $\varphi(\mathbf{x})$ *is convex and the feasible region is determined by linear*

constraints:

$$X = \{\mathbf{x}: \langle \mathbf{a}_i, \mathbf{x}\rangle - b_i \geqslant 0, \quad i = \overline{1, m}, \mathbf{x} \geqslant 0\},$$

then for a point $\mathbf{x}^* \in X$ *to be optimal it is necessary and sufficient that there exist* $\mathbf{y}^* > 0$ *such that the pair* $\mathbf{x}^*$, $\mathbf{y}^*$ *is a saddle point of the Lagrangian function on the set* $\mathbf{x} \geqslant 0$, $\mathbf{y} \geqslant 0$.

The theorem follows from 3.5.4 and 3.6.3. △

3.6.10. There are different versions of the Kuhn-Tucker theorem and its various generalizations and applications in the theory of extremal problems. This theorem lies in the foundation of the theory of duality in mathematical programming.

The Kuhn-Tucker theorem is also used in numerical methods of solving mathematical programming problems. It makes it possible to replace the original problem by the problem of finding a saddle point of the Lagrangian function, i.e. a problem having the form

$$\min_{\mathbf{x}\in\Gamma} \max_{\mathbf{y}\geqslant 0} L(\mathbf{x}, \mathbf{y}).$$

The "simple" constraints in the latter problem allow one to solve it using methods which are in many respects similar to numerical methods of unconditional optimization, which are rather thoroughly investigated and evaluated.

Remark. The Kuhn-Tucker theorem becomes meaningless in the case of degeneracy, for instance, when

$$\varphi'(\mathbf{x}^*) \equiv 0, \quad f_i'(\mathbf{x}^*) \equiv 0, \quad i = \overline{1, m}.$$

However, there is a wide class of such problems. In Supplement 3 we present necessary and sufficient optimality conditions for degenerate problems.

3.6.11. Duality. We introduce the function $g(\mathbf{x}) = \sup_{\mathbf{y}\geqslant 0} L(\mathbf{x}, \mathbf{y})$. It is obvious that the basic mathematical programming problem can be written as

$$g(\mathbf{x}) \to \min, \quad \mathbf{x} \in X = \{\mathbf{x}: \mathbf{f}(\mathbf{x}) \geqslant 0\},$$

since $g(\mathbf{x}) = \varphi(\mathbf{x})$ for $\mathbf{x} \in X$. This problem is usually called *primal*.

Denote $h(\mathbf{y}) = \inf\limits_{\mathbf{x}\in E_n} L(\mathbf{x}, \mathbf{y})$. The problem

$$h(\mathbf{y}) \to \max, \quad \mathbf{y} \in E_m^+ = \{\mathbf{y}: \mathbf{y} \geqslant \mathbf{0}\},$$

is called *dual* and $y_1, y_2, \ldots, y_m$ are called *dual variables.*

3.6.12. Theorem. (1) $\varphi(\mathbf{x}) \geqslant h(\mathbf{y})$ *for all* $\mathbf{x} \in X$, $\mathbf{y} \in E_m^+$.

(2) *If the conditions of the Kuhn-Tucker theorem* (Sec. 3.6.7) *hold and the pair* $\mathbf{x}^*$, $\mathbf{y}^*$ *is a saddle point of the Lagrangian function, then*

$$\mathbf{y}^* = \arg\max\{h(\mathbf{y}): \mathbf{y} \in E_m^+\} \quad \text{and} \quad \varphi(\mathbf{x}^*) = h(\mathbf{y}^*).$$

(3) *If* $\varphi(\mathbf{x}^*) = h(\mathbf{y}^*)$ *for any* $\mathbf{x}^* \in X$, $\mathbf{y}^* \in E_m^+$, *then*

$$\mathbf{x}^* = \arg\min\{\varphi(\mathbf{x}): \mathbf{x} \in X\},$$
$$\mathbf{y}^* = \arg\max\{h(\mathbf{y}): \mathbf{y} \in E_m^+\}.$$

Proof. (1) Since $\mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}$, $\mathbf{y} \geqslant \mathbf{0}$, we have

$$\varphi(\mathbf{x}) \geqslant \varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x})\rangle = L(\mathbf{x}, \mathbf{y}) \geqslant \inf_{\mathbf{x}\in E_n} L(\mathbf{x}, \mathbf{y}) = h(\mathbf{y}).$$

(2) The inequality

$$h(\mathbf{y}^*) = \inf_{\mathbf{x}\in E_n} L(\mathbf{x}, \mathbf{y}^*) = L(\mathbf{x}^*, \mathbf{y}^*) \geqslant L(\mathbf{x}^*, \mathbf{y}) \geqslant \inf_{\mathbf{x}\in E_n} L(\mathbf{x}, \mathbf{y}) = h(\mathbf{y})$$

if fulfilled for all $\mathbf{y} \geqslant \mathbf{0}$, and therefore $\mathbf{y}^*$ is a solution to the dual problem. However, $L(\mathbf{x}^*, \mathbf{y}^*) = \varphi(\mathbf{x}^*)$ since $\langle \mathbf{y}^*, f(\mathbf{x}^*)\rangle = 0$ (see Sec. 3.31), and hence $\varphi(\mathbf{x}^*) = h(\mathbf{y}^*)$.

(3) By virtue of (1), $\varphi(\mathbf{x}) \geqslant h(\mathbf{y}^*) = \varphi(\mathbf{x}^*) \geqslant h(\mathbf{y})$ for any $\mathbf{x} \in X$, $\mathbf{y} \in E_m^+$, and, consequently, $\mathbf{x}^*$ is a solution to the primal problem and $\mathbf{y}^*$ is a solution to the dual problem.

Definition. A point $\mathbf{v}^* = (\mathbf{x}^*, \mathbf{y}^*)$ satisfying the conditions

$$\varphi'(\mathbf{x}^*) - \sum_{i=1}^{m} y_i^* \varphi'(\mathbf{x}^*) = \mathbf{0}, \quad y_i^* \geqslant 0,$$
$$y_i^* f_i(\mathbf{x}^*) = 0, \quad i = \overline{1, m}$$

will be called a *Kuhn-Tucker point.*

Chapter 4

LINEAR PROGRAMMING THEORY

4.1. Basic Notions

4.1.1. Basic problem. The following extremal problem will be called the *basic linear programming problem*:

$$\min_{x \in R_1} \langle \mathbf{c}, \mathbf{x} \rangle,$$

$$R_1 = \{\mathbf{x}: A\mathbf{x} \geqslant \mathbf{b}, \mathbf{x} \geqslant 0\}. \tag{4.1}$$

It will be shown later that any linear programming problem can be reduced to the form (4.1). Therefore the fundamental linear programming theorems will be considered in application to the problem (4.1), which by no means leads to a loss of generality.

4.1.2. Duality. By the problem dual to the basic linear programming problem will be meant

$$\max_{y \in Q_1} \langle \mathbf{b}, \mathbf{y} \rangle,$$

$$Q_1 = \{\mathbf{y}: A^T\mathbf{y} \leqslant \mathbf{c}, \mathbf{y} \geqslant 0\}. \tag{4.2}$$

4.1.3. Equivalence. Two extremal problems will be called *equivalent* if either their solution sets coincide or both of the problems are unsolvable.

It can easily be seen that the problem (4.1) is dual to (4.2). To this end it suffices to consider the problem

$$\min_{y \in Q_1} \langle -\mathbf{b}, \mathbf{y} \rangle,$$

$$Q_1 = \{\mathbf{y}: -A^T\mathbf{y} \geqslant -\mathbf{c}, \mathbf{y} \geqslant 0\}$$

equivalent to (4.2) and to construct, in accordance with the definition, its dual problem:

$$\max_{x \in R_1} \langle -\mathbf{c}, \mathbf{x} \rangle,$$

$$R_1 = \{\mathbf{x}: -A\mathbf{x} \leqslant -\mathbf{b}, \mathbf{x} \geqslant 0\}.$$

However, this is the problem (4.1).

Thus, the problems (4.1) and (4.2) are mutually dual.

4.1.4. Terminology. Throughout this book we shall use the earlier introduced terminology (an objective function, a feasible region, a feasible point, an optimal point, etc.).

However, it should be noted that, generally, the terminology in linear programming is somewhat ambiguous. This is to a great extent due to economic applications

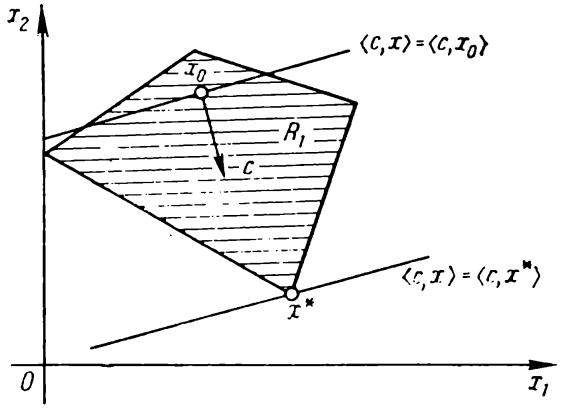


Fig. 4.1

whose terminology is often retained in linear programming problems as well.

The matrix $A = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n]$ is called the *condition matrix* of the problem (4.1).

Its columns $\mathbf{a}_i$ $(i = \overline{1, n})$ are called *condition vectors* of the problem (4.1).

The vector $\mathbf{b}$ is called the *constraint vector* of the problem (4.1).

4.1.5. Geometrical interpretation. The set R_1 can be regarded in the space E_n as an intersection of half-spaces (half-planes for $n = 2$):

$$(A\mathbf{x})_i \geqslant b_i, \quad i = \overline{1, m},$$
$$x_j \geqslant 0, \quad j = \overline{1, n}.$$

Consider the family

$$\langle \mathbf{c}, \mathbf{x} \rangle = \lambda$$

of parallel hyperplanes (parallel straight lines for $n = 2$). The vector $-\mathbf{c}$ goes in the direction of decrease of the objective function. Figure 4.1 represents the case when the set R_1 is bounded (it is a polygon).

Consider a point $\mathbf{x}_0 \in R_1$. The corresponding value of the objective function is

$$\lambda_0 = \langle \mathbf{c}, \mathbf{x}_0 \rangle.$$

We now begin to translate the line

$$\langle \mathbf{c}, \mathbf{x} \rangle = \lambda$$

in the direction $-\mathbf{c}$, i.e. in the direction of decrease of λ, until we arrive at a point $\mathbf{x}^* \in R_1$ where λ attains minimum. Here the geometrical meaning of the problem is obvious.

It is also easy to represent the cases when R_1 is unbounded but the solution exists; when the solution exists but is not unique; and, finally, when R_1 is unbounded and $\langle \mathbf{c}, \mathbf{x} \rangle$ is unbounded on R_1.

4.2. Basic Theorems

4.2.1. We first state a theorem which is a special case of Theorem 3.6.9 (when $\varphi(\mathbf{x})$ is linear) and therefore needs no proof.

The Lagrangian function for the problem (4.1) will be denoted thus:

$$L_1(\mathbf{x}, \mathbf{y}) = \langle \mathbf{c}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x} \rangle. \tag{4.3}$$

Theorem. *For $\mathbf{x}^* \in R_1$ to be an optimal point for the basic linear programming problem it is necessary and sufficient that there exist $\mathbf{y}^* \geqslant \mathbf{0}$ such that the pair $\mathbf{x}^*$, $\mathbf{y}^*$ is a saddle point of the Lagrangian function $L_1(\mathbf{x}, \mathbf{y})$ in the region $\mathbf{x} \geqslant \mathbf{0}$, $\mathbf{y} \geqslant \mathbf{0}$, i.e.*

$$L_1(\mathbf{x}^*, \mathbf{y}) \leqslant L_1(\mathbf{x}^*, \mathbf{y}^*) \leqslant L_1(\mathbf{x}, \mathbf{y}^*). \tag{4.4}$$

4.2.2. As is implied by (3.33), the following representation of the vector $\mathbf{c}$ is a necessary and sufficient condition for the optimality of $\mathbf{x}^*$ in the problem (4.1):

$$-\mathbf{c} = -\sum_{i \in I(\mathbf{x}^*)} y_i^* \mathbf{a}_i - \sum_{j \in J(\mathbf{x}^*)} v_i^* \mathbf{e}_j, \quad y_i^* \geqslant 0, \quad v_j^* \geqslant 0, \tag{4.5}$$

where

$$I(\mathbf{x}^*) = \{i\colon (A\mathbf{x}^*)_i = b_i\}, \quad J(\mathbf{x}^*) = \{j\colon x_j^* = 0\}.$$

From Theorems 3.6.2 and 4.2.1 it follows that $\mathbf{x}^*$ and $\mathbf{y}^*$ satisfy

$$\begin{aligned} x_i^* (\mathbf{c} - A^T\mathbf{y}^*)_i &= 0, \quad i = \overline{1, n}, \\ y_j^* (\mathbf{b} - A\mathbf{x}^*)_j &= 0, \quad j = \overline{1, m}. \end{aligned} \tag{4.6}$$

4.2.3. Duality theorems. The first duality theorem. *Both the primal and dual problems either have optimal points* $\mathbf{x}^*$ *and* $\mathbf{y}^*$, *which must satisfy*

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{b}, \mathbf{y}^* \rangle, \tag{4.7}$$

or do not have them.

Proof. Consider the Lagrangian function for the problem (4.2). To this end we write the problem in the form of the basic linear programming problem, i.e. as (4.1).

It is clear that the problem

$$\max_{\mathbf{y}\in Q_1} \langle \mathbf{b}, \mathbf{y} \rangle,$$
$$Q_1 = \{\mathbf{y}\colon A^T\mathbf{y} \leqslant \mathbf{c}, \quad \mathbf{y} \geqslant \mathbf{0}\}$$

is equivalent to the problem

$$\min_{\mathbf{y}\in Q_1} \langle -\mathbf{b}, \mathbf{y} \rangle,$$
$$Q_1 = \{\mathbf{y}\colon -A^T\mathbf{y} \geqslant -\mathbf{c}, \quad \mathbf{y} \geqslant \mathbf{0}\}.$$

Since the latter is written as the basic problem, its Lagrangian function is

$$L_2(\mathbf{y}, \mathbf{x}) = -\langle \mathbf{b}, \mathbf{y} \rangle + \langle \mathbf{x}, -\mathbf{c} + A^T\mathbf{y} \rangle. \tag{4.8}$$

A saddle point for $L_2(\mathbf{y}, \mathbf{x})$ in the region $\mathbf{y} \geqslant \mathbf{0}$, $\mathbf{x} \geqslant \mathbf{0}$ is a pair $\mathbf{y}'$, $\mathbf{x}'$ such that

$$L_2(\mathbf{y}', \mathbf{x}) \leqslant L_2(\mathbf{y}', \mathbf{x}') \leqslant L_2(\mathbf{y}, \mathbf{x}'). \tag{4.9}$$

Comparing (4.8) and (4.3) we find

$$\begin{aligned} L_2(\mathbf{y}, \mathbf{x}) &= -[\langle \mathbf{c}, \mathbf{x} \rangle + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x} \rangle] \\ &= -L_1(\mathbf{x}, \mathbf{y}). \end{aligned} \tag{4.10}$$

It follows from (4.10), (4.9), and (4.4) that if $\mathbf{x}^*$, $\mathbf{y}^*$ is a saddle point for $L_1(\mathbf{x}, \mathbf{y})$, then $\mathbf{y}^*$, $\mathbf{x}^*$ is a saddle point for $L_2(\mathbf{y}, \mathbf{x})$, and, consequently, either $\mathbf{x}^*$ and $\mathbf{y}^*$ are

optimal points for the problems (4.1) and (4.2), respectively, or, when there is no saddle point, both problems (4.1) and (4.2) have no solutions.

Finally, (4.7) follows from the formulas (4.6). △

Definitions. We shall call the conditions

$$(A\mathbf{x})_j \geqslant b_j, \quad y_j \geqslant 0, \quad j = \overline{1, m},$$

in the problems (4.1) and (4.2) a *pair of dual conditions.*

Similarly,

$$x_j \geqslant 0, \quad (A^T\mathbf{y})_i \leqslant c_i, \quad i = \overline{1, n},$$

are also a pair of dual conditions.

Any of the conditions is called *free* if it is fulfilled as a strict inequality for at least one optimal vector.

A condition is called *fixed* if it is fulfilled as an equation for all optimal vectors.

The second duality theorem. *If the primal problem (and, hence, the dual problem as well) is solvable, then in each pair of dual conditions one condition is free and the other is fixed.*

Proof. The assertion that the dual condition corresponding to a free condition is fixed follows immediately from (4.6). Let us show that the dual condition corresponding to a fixed condition is free. For instance, let $x_l \geqslant 0$ be a fixed condition: $x_l^* = 0$ for all $\mathbf{x}^* \in X^*$ where X^* is the set of solutions to the problem (4.1). We construct a solution $\bar{\mathbf{y}}$ to the problem (4.2) for which $(A^T\bar{\mathbf{y}})_l < c_l$. We shall use induction on the dimension n.

For $n = 1$ we have

$$\begin{aligned} &\min \langle c_1, x_1 \rangle, \quad c_1 \neq 0, \\ &Ax_1 \geqslant \mathbf{b}, \quad x_1 \geqslant 0 \\ &(A \text{ is a one-column matrix}) \end{aligned}$$

in the primal problem; if $x_1^* = 0$, then $c_1 > 0$ in the dual problem:

$$\begin{aligned} &\max \langle \mathbf{b}, \mathbf{y} \rangle, \\ &A^T\mathbf{y} \leqslant c_1, \\ &\mathbf{y} \geqslant \mathbf{0}, \end{aligned}$$

$\bar{\mathbf{y}} = \mathbf{0}$ is a solution, and $(A^T\bar{\mathbf{y}})_1 \equiv A^T\bar{\mathbf{y}} = 0 < c_1$.

Assume that the theorem is true for $n-1$. We construct $\bar{\mathbf{y}}$ for the dimension n. Consider the auxiliary problem:

$$\min_{\mathbf{x}\in R_1(\varepsilon)} \langle \mathbf{c}, \mathbf{x}\rangle, \qquad R_1(\varepsilon)=\{\mathbf{x}:\ A\mathbf{x}\geqslant\mathbf{b},\ \mathbf{x}\geqslant\mathbf{0},\ x_l\geqslant\varepsilon>0\}. \tag{A}$$

(1) If $R_1(\varepsilon)=\varnothing$ for all $\varepsilon>0$, then $x_l=0$ for all $x\in R_1$, and, consequently, the problem (4.1) has the dimension $n-1$ for which the theorem is true by the induction hypothesis.

(2) Let $R_1(\varepsilon)\neq\varnothing$ for all sufficiently small $\varepsilon>0$. Since the condition $x_l\geqslant 0$ is fixed and $R_1(\varepsilon)\subseteq R_1=\{\mathbf{x}:\ A\mathbf{x}\geqslant\mathbf{b},\ \mathbf{x}\geqslant\mathbf{0}\}$, we have

$$\min_{\mathbf{x}\in R_1(\varepsilon)}\langle\mathbf{c},\mathbf{x}\rangle=\varphi_\varepsilon^*>\varphi^*=\min_{\mathbf{x}\in R_1}\langle\mathbf{c},\mathbf{x}\rangle$$

(if we assume that $\varphi_\varepsilon^*=\varphi^*$, then there is $x_l^*\geqslant\varepsilon>0$, which contradicts the fact that the condition is fixed).

We construct the problem dual to (A):

$$\max_{\mathbf{y}\in Q}(\langle\mathbf{b},\mathbf{y}\rangle+\varepsilon y_{m+1}), \qquad Q=\{\mathbf{y}:\ A^T\mathbf{y}+e_l y_{m+1}\leqslant\mathbf{c},\ \mathbf{y}\geqslant 0,\ y_{m+1}\geqslant 0\}. \tag{B}$$

Let a point $\begin{bmatrix}\bar{\mathbf{y}}(\varepsilon)\\ \bar{y}_{m+1}(\varepsilon)\end{bmatrix}$ be a solution to the problem (B). We shall show that the point $\bar{\mathbf{y}}(\varepsilon)$ is a solution to the problem (4.2). Without loss of generality, we assume that the active constraints are

$$(A^T\bar{\mathbf{y}}(\varepsilon))_i=c_i,\quad i=\overline{1,k},\quad i\neq l,\quad k\leqslant n,$$
$$(A^T\bar{\mathbf{y}}(\varepsilon))_l+\bar{y}_{m+1}(\varepsilon)=c_l,$$
$$\bar{y}_j(\varepsilon)=0,\quad j=\overline{1,p},\quad j\neq m+1,\quad p\leqslant m.$$

By the Kuhn-Tucker theorem (3.5.1), in the solution $\begin{bmatrix}\bar{\mathbf{y}}(\varepsilon)\\ \bar{y}_{m+1}(\varepsilon)\end{bmatrix}$ there must be

$$\begin{bmatrix}\mathbf{b}\\ \varepsilon\end{bmatrix}=\sum_{\substack{i=\overline{1,k}\\ i\neq l}}\lambda_i\begin{bmatrix}\mathbf{a}_i\\ 0\end{bmatrix}+\lambda_l\begin{bmatrix}\mathbf{a}_l\\ 1\end{bmatrix}-\sum_{j=\overline{1,p}}u_j\begin{bmatrix}\mathbf{e}_j\\ 0\end{bmatrix},$$

$$\lambda_i\geqslant 0,\quad i=\overline{1,k},\quad i\neq l,\quad \lambda_l>0,\quad u_j\geqslant 0,\quad j=\overline{1,p}.$$

For $\varepsilon \to +0$ we have $\lambda_l \to +0$. Therefore, by Theorem (2.3.2), there are nonnegative numbers

$$\lambda_i^* \geqslant 0 \ (i = \overline{1, k}, \ i \neq l), \quad \lambda_l^*, \ u_j^* \geqslant 0 \ (j = \overline{1, p}, \ j \neq m+1),$$

such that

$$\begin{bmatrix} \mathbf{b} \\ 0 \end{bmatrix} = \sum_{\substack{i=\overline{1,k} \\ i \neq l}} \lambda_i^* \begin{bmatrix} \mathbf{a}_i \\ 0 \end{bmatrix} + \lambda_l^* \begin{bmatrix} \mathbf{a}_l \\ 1 \end{bmatrix} - \sum_{j=\overline{1,p}} u_j^* \begin{bmatrix} \mathbf{e}_j \\ 0 \end{bmatrix}, \qquad \text{(C)}$$

where $\lambda_l^* = 0$. Further, at the points $\bar{\mathbf{y}}(\varepsilon)$ we have $(A^T \bar{\mathbf{y}}(\varepsilon))_l + \bar{y}_{m+1}(\varepsilon) = c_l$, and, since $\bar{y}_{m+1}(\varepsilon) > 0$, there holds $(A^T \bar{\mathbf{y}}(\varepsilon))_l < c_l$. Therefore only the constraints

$$(A^T \bar{\mathbf{y}}(\varepsilon))_i = c_i, \quad i = \overline{1, k}, \quad i \neq l, \quad \bar{y}_j(\varepsilon) = 0, \quad j = \overline{1, p},$$

are active in the problem (4.2) at the points $\bar{\mathbf{y}}(\varepsilon)$. Rewriting (C) as

$$\mathbf{b} = \sum_{\substack{i=\overline{1,k} \\ i \neq l}} \lambda_i^* \mathbf{a}_i - \sum_{j=\overline{1,p}} u_j^* \mathbf{e}_j$$

and taking the above remark into account we immediately arrive at the conditions of Theorem (3.5.1), which implies that $\bar{\mathbf{y}}(\varepsilon)$ is a solution to the problem (4.2) for sufficiently small $\varepsilon > 0$. △

4.2.4. Theorem. *For any feasible* $\mathbf{x} \in R_1$ *and* $\mathbf{y} \in Q_1$ *the inequality*

$$\langle \mathbf{c}, \ \mathbf{x} \rangle \geqslant \langle \mathbf{b}, \ \mathbf{y} \rangle \qquad (4.11)$$

is fulfilled.

Proof. The inequalities determining R_1 and Q_1 imply

$$\langle \mathbf{c}, \ \mathbf{x} \rangle \geqslant \langle A^T \mathbf{y}, \ \mathbf{x} \rangle = \langle \mathbf{y}, \ A\mathbf{x} \rangle \geqslant \langle \mathbf{y}, \ \mathbf{b} \rangle = \langle \mathbf{b}, \ \mathbf{y} \rangle. \ \triangle$$

4.2.5. Theorem. *If* $\mathbf{x}^* \in R_1$, $\mathbf{y}^* \in Q_1$, *and*

$$\langle \mathbf{c}, \ \mathbf{x}^* \rangle = \langle \mathbf{b}, \ \mathbf{y}^* \rangle,$$

then $\mathbf{x}^*$ *and* $\mathbf{y}^*$ *are optimal for the problems* (4.1) *and* (4.2), *respectively, and vice versa.*

Proof. By virtue of (4.11) and the statement of the theorem, for any $\mathbf{x} \in R_1$ we have

$$\langle \mathbf{c}, \ \mathbf{x} \rangle \geqslant \langle \mathbf{b}, \ \mathbf{y}^* \rangle = \langle \mathbf{c}, \ \mathbf{x}^* \rangle,$$

i.e. $\mathbf{x}^*$ is optimal. The optimality of $\mathbf{y}^*$ is proved in a similar way. The converse assertion in the theorem is implied by the duality theorem 4.2.3. $\triangle$

4.2.6. Theorem. *If* $\inf\limits_{\mathbf{x}\in R_1} \langle \mathbf{c}, \mathbf{x}\rangle = M > -\infty$, *then there exists a point*

$$\mathbf{x}^* = \arg\min \{\langle \mathbf{c}, \mathbf{x}\rangle : \mathbf{x} \in R_1\}.$$

Proof. Assume that there is no such point. Then the system of inequalities

$$\begin{gathered} A\mathbf{x} \geqslant \mathbf{b}, \\ \langle \mathbf{c}, \mathbf{x}\rangle \leqslant M, \\ \mathbf{x} \geqslant \mathbf{0} \end{gathered} \tag{4.12}$$

has no solution. Let

$$B = \begin{bmatrix} -A \\ \mathbf{c}^T \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} -\mathbf{b} \\ M \end{bmatrix}.$$

We rewrite the system (4.12) in the form

$$\begin{gathered} B\mathbf{x} \leqslant \mathbf{d}, \\ \mathbf{x} \geqslant \mathbf{0}. \end{gathered}$$

Since this system has no solution, the modification of the Farkas theorem (see 2.3.5) implies the solvability of the following system:

$$B^T\mathbf{u} \geqslant \mathbf{0}, \tag{4.13}$$

$$\langle \mathbf{d}, \mathbf{u}\rangle < 0, \tag{4.14}$$

$$\mathbf{u} \geqslant \mathbf{0}. \tag{4.15}$$

If the number M in the inequality (4.14)

$$-\sum_{i=1}^{m} b_i u_i + M u_{m+1} < 0$$

is given a sufficiently small increment $\varepsilon > 0$, the system

$$\begin{gathered} B^T\mathbf{u} \geqslant \mathbf{0}, \\ -\sum_{i=1}^{m} b_i u_i + (M+\varepsilon)\, u_{m+1} < 0, \\ \mathbf{u} \geqslant \mathbf{0} \end{gathered}$$

becomes solvable. In view of the modification of the Farkas theorem, the system

$$B\mathbf{x} \leqslant \mathbf{d}_\varepsilon,$$
$$\mathbf{x} \geqslant \mathbf{0},$$

where $\mathbf{d}_\varepsilon = \begin{bmatrix} -\mathbf{b} \\ M+\varepsilon \end{bmatrix}$, has no solution. In the old notation this system has the form

$$A\mathbf{x} \geqslant \mathbf{b},$$
$$\langle \mathbf{c}, \mathbf{x} \rangle \leqslant M + \varepsilon,$$
$$\mathbf{x} \geqslant \mathbf{0}.$$

However, the assertion that this system is unsolvable contradicts the definition of the number M as the infimum of the function $\langle \mathbf{c}, \mathbf{x} \rangle$ on the set R_1. $\triangle$

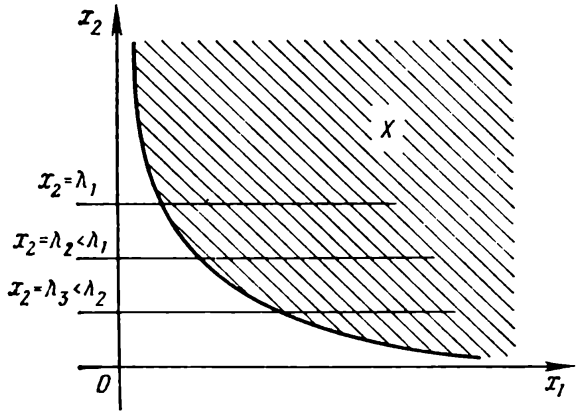


Fig. 4.2

4.2.7. It should be noted that, generally, the theorem is true only for linear programming problems.

For example, let us consider the following problem: minimize x_2 on the set

$$X = \left\{ \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} : \ x_1 \cdot x_2 \geqslant 1, \quad x_1 \geqslant 0, \quad x_2 \geqslant 0 \right\}.$$

It is obvious that the set X is convex and closed and

$$\inf_{x_2 \in X} x_2 = 0;$$

however, X contains no point where $\inf_X x_2$ is attained (Fig. 4.2).

4.2.8. Theorem. *For a solution of one of the dual problems (and, consequently, of both problems) to exist it is necessary and sufficient that*

$$R_1 = \varnothing \quad \text{and} \quad Q_1 \neq \varnothing.$$

Proof. The necessity is obvious.

Sufficiency. Let $\mathbf{y} \in Q_1$; then for any $\mathbf{x} \in R_1$ we have (4.11):

$$\langle \mathbf{c}, \mathbf{x} \rangle \geqslant \langle \mathbf{b}, \mathbf{y} \rangle,$$

and therefore, by Theorem 4.2.6, the problem (4.1) has a solution and hence the problem (4.2) also has a solution (by virtue of Theorem 4.2.3). △

4.2.9. *If** $\langle \mathbf{c}, \mathbf{x} \rangle$ ($\langle \mathbf{b}, \mathbf{y} \rangle$) *is unbounded from below on* R_1 *(from above on* Q_1*), then*

$$Q_1 = \varnothing \quad (R_1 = \varnothing).$$

4.2.10. *If* $Q_1 = \varnothing$ ($R_1 = \varnothing$) *and* $R_1 \neq \varnothing$ ($Q_1 \neq \varnothing$), *then* R_1 (Q_1) *is unbounded and* $\langle \mathbf{c}, \mathbf{x} \rangle$ *is unbounded from below on* R_1 ($\langle \mathbf{b}, \mathbf{y} \rangle$ *is unbounded from above on* Q_1).

4.2.11. Theorem. *If*

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \min_{R_1} \langle \mathbf{c}, \mathbf{x} \rangle$$

and if there exist $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ *such that*

(a) $\mathbf{x}_i \in R_1$, $i = \overline{1, M}$;

(b) $\mathbf{x}^* = \sum_{i=1}^{M} \alpha_i \mathbf{x}_i$, *where*

$$\sum_{i=1}^{M} \alpha_i = 1 \quad (\alpha_i > 0,\ i = \overline{1, M}),$$

then

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_1 \rangle = \ldots = \langle \mathbf{c}, \mathbf{x}_M \rangle.$$

Proof. Assume that

$$\langle \mathbf{c}, \mathbf{x}^* \rangle \leqslant \langle \mathbf{c}, \mathbf{x}_1 \rangle \leqslant \ldots \leqslant \langle \mathbf{c}, \mathbf{x}_M \rangle. \tag{4.16}$$

Then

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \left\langle \mathbf{c}, \sum_{i=1}^{M} \alpha_i \mathbf{x}_i \right\rangle = \sum_{i=1}^{M} \alpha_i \langle \mathbf{c}, \mathbf{x}_i \rangle \geqslant \sum_{i=1}^{M} \alpha_i \langle \mathbf{c}, \mathbf{x}_1 \rangle = \langle \mathbf{c}, \mathbf{x}_1 \rangle.$$

* Given in the parentheses is the dual assertion.

Taking (4.16) into account we find

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_1 \rangle.$$

We state the induction hypothesis:

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_1 \rangle = \ldots = \langle \mathbf{c}, \mathbf{x}_k \rangle, \quad k < M,$$

and prove that

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle.$$

Indeed,

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \sum_{i=1}^{M} \alpha_i \langle \mathbf{c}, \mathbf{x}_i \rangle = \sum_{i=1}^{k} \alpha_i \langle \mathbf{c}, \mathbf{x}_i \rangle + \sum_{i=k+1}^{M} \alpha_i \langle \mathbf{c}, \mathbf{x}_i \rangle$$

$$\geqslant \langle \mathbf{c}, \mathbf{x}^* \rangle \sum_{i=1}^{k} \alpha_i + \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle \sum_{i=k+1}^{M} \alpha_i,$$

whence

$$\Big(1 - \sum_{i=1}^{k} \alpha_i\Big) \langle \mathbf{c}, \mathbf{x}^* \rangle \geqslant \Big(\sum_{i=k+1}^{M} \alpha_i\Big) \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle$$

$$= \Big(1 - \sum_{i=1}^{k} \alpha_i\Big) \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle.$$

Further, since $\sum_{i=1}^{k} \alpha_i < 1$ (because $k < M$) we have

$$\langle \mathbf{c}, \mathbf{x}^* \rangle \geqslant \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle,$$

which implies (see (4.16))

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_{k+1} \rangle. \quad \triangle$$

4.2.12. Theorem. *If the problem* (4.1) *possesses a solution* $\mathbf{x}^*$, *then there is an extreme point* $\bar{\mathbf{x}}$ *such that*

$$\langle \mathbf{c}, \bar{\mathbf{x}} \rangle = \min_{\mathbf{x} \in R_1} \langle \mathbf{c}, \mathbf{x} \rangle.$$

Proof. (1) If R_1 is bounded, then Theorem 2.2.10 implies that there are extreme points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ of the set R_1 such that for any $\mathbf{x} \in R_1$ and, in particular, for $\mathbf{x}^*$ we have

$$\mathbf{x}^* = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i, \quad \sum_{i=1}^{N} \alpha_i = 1, \quad \alpha_i \geqslant 0, \quad i = \overline{1, N}.$$

Delete from the sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ those points to which $\alpha_i = 0$ correspond. Without loss of generality, we

can assume these are the points $\mathbf{x}_{M+1}, \mathbf{x}_{M+2}, \ldots, \mathbf{x}_N$. Then

$$\mathbf{x}^* = \sum_{i=1}^{M} \alpha_i \mathbf{x}_i, \quad \sum_{i=1}^{M} \alpha_i = 1, \quad \alpha_i > 0, \quad i = \overline{1, M},$$

i.e. the conditions of Theorem 4.2.11 are fulfilled, whence follows the assertion of the theorem.

(2) Let R_1 be unbounded and let $\mathbf{x}^*$ be optimal. It is obvious that for a sufficiently large number $\mu > 0$ the set

$$L = \left\{\mathbf{x}: \sum_{i=1}^{n} x_i \leqslant \mu, \quad x_i \geqslant 0, \quad i = \overline{1, n}\right\}$$

is such that

$$\mathbf{x}^* \in R_1 \cap L, \text{ and } \mathbf{x}^* \notin l = \left\{\mathbf{x}: \sum_{i=1}^{n} x_i = \mu, \ x_i \geqslant 0, \ i = \overline{1, n}\right\}.$$

Since the intersection of the "orthant" $\mathbf{x} \geqslant \mathbf{0}$ with the set L is bounded, the set $R_1 \cap L$ is also bounded, and, consequently (by 2.2.10 and 4.2.11), there are extreme points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M$ of $R_1 \cap L$ such that

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{c}, \mathbf{x}_1 \rangle = \ldots = \langle \mathbf{c}, \mathbf{x}_M \rangle.$$

If at least one of the points $\mathbf{x}_i$ does not belong to l, then the theorem is proved since in this case $\mathbf{x}_i$ is an extreme point of the set R_1. If all $\mathbf{x}_i$ belong to l ($i = \overline{1, M}$), then the representation

$$\mathbf{x}^* = \sum_{i=1}^{M} \alpha_i \mathbf{x}_i, \quad \sum_{i=1}^{M} \alpha_i = 1, \quad \alpha_i > 0, \quad i = \overline{1, M},$$

implies $\mathbf{x}^* \in l$, which contradicts the choice of the number $\mu > 0$ in the construction of the set L. $\triangle$

4.3. Algebraic Characterization of an Extreme Point

4.3.1. Earlier we stated a geometrical definition of an extreme point. However, in order to be able to find (calculate) an extreme point it is necessary to characterize

it algebraically. As before, let

$$I(\mathbf{x}) = \{i\colon (A\mathbf{x})_i = b_i\}, \quad J(\mathbf{x}) = \{j\colon x_j = 0\},$$
$$J = \{j = 1, 2, \ldots, n\}.$$

Consider the system of equations

$$(A\mathbf{z})_i = b_i, \quad i = I(\mathbf{x}), \quad z_j = 0, \quad j \in J(\mathbf{x}). \tag{4.17}$$

Without loss of generality, we can assume that

$$I(\mathbf{x}) = \{i\colon i = 1, 2, \ldots, r\},$$
$$J(\mathbf{x}) = \{j\colon j = k+1, k+2, \ldots, n\}.$$

Then, for $r = k$, the system (4.17) is square.

4.3.2. Theorem. *For* $\mathbf{x} \neq \mathbf{0}$ *to be an extreme point of the set* R_1 *it is necessary and sufficient that the vector* $\mathbf{x}$ *satisfy the non-singular square system of equations* (4.17).

Proof. Sufficiency. Let $\mathbf{x} \in R_1$ and let there exist a matrix

$$B = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1k} \\ a_{21} & a_{22} & \ldots & a_{2k} \\ \ldots & \ldots & \ldots & \ldots \\ a_{k1} & a_{k2} & \ldots & a_{kk} \end{bmatrix}, \quad \det B \neq 0, \tag{4.18}$$

such that

$$B\bar{\mathbf{x}} = \bar{\mathbf{b}}. \tag{4.19}$$

Here

$$\bar{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \ldots \\ x_k \end{bmatrix}, \quad \bar{\mathbf{b}} = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_k \end{bmatrix}, \quad x_{k+1} = x_{k+2} = \ldots = x_n = 0,$$

$$\mathbf{x} = \begin{bmatrix} \bar{\mathbf{x}} \\ 0 \\ \ldots \\ 0 \end{bmatrix}. \tag{4.20}$$

Assume that $\mathbf{x}$ is not an extreme point, i.e. there are $\mathbf{x}' \in R_1$, $\mathbf{x}'' \in R_1$, $\mathbf{x}' \neq \mathbf{x}'' \neq \mathbf{x}$, such that

$$\mathbf{x} = \alpha\mathbf{x}' + (1 - \alpha)\,\mathbf{x}'', \quad \alpha \in (0, 1).$$

For $j > k$ we have

$$x_j = \alpha x'_j + (1 - \alpha)\, x''_j = 0,$$

and, since $\alpha > 0$, $(1 - \alpha) > 0$, $x'_j \geqslant 0$, and $x''_j \geqslant 0$,

$$x'_j = x''_j = 0, \quad j = \overline{k+1,\ n}.$$

Further,

$$B\bar{\mathbf{x}}' \geqslant \bar{\mathbf{b}}, \quad B\bar{\mathbf{x}}'' \geqslant \bar{\mathbf{b}}, \quad \alpha B\bar{\mathbf{x}}' + (1-\alpha)\, B\bar{\mathbf{x}}'' = \bar{\mathbf{b}},$$

and, consequently,

$$B\bar{\mathbf{x}}' = B\bar{\mathbf{x}}'' = \bar{\mathbf{b}}.$$

Therefore, since $\det B \neq 0$, we have $\bar{\mathbf{x}}' = \bar{\mathbf{x}}''$; hence, $\mathbf{x}' = \mathbf{x}''$, which contradicts the assumption.

Necessity. Let $\mathbf{x}$ be an extreme point of the set R_1. (1) We shall show that, for at least one i,

$$(A\mathbf{x})_i = b_i.$$

Assume that there is no such index i. Since $\mathbf{x} \neq \mathbf{0}$, there is j such that $x_j > 0$. Consider

$$(\mathbf{x}')^T = (x_1,\ x_2,\ \ldots,\ x_{j-1},\ x_j + \varepsilon,\ x_{j+1},\ \ldots,\ x_n) \geqslant \mathbf{0}$$

and

$$(\mathbf{x}'')^T = (x_1,\ x_2,\ \ldots,\ x_{j-1},\ x_j - \varepsilon,\ x_{j+1},\ \ldots,\ x_n) \geqslant \mathbf{0}.$$

The assumption that $A\mathbf{x} > \mathbf{b}$ implies that for a sufficiently small ε we have

$$A\mathbf{x}' \geqslant \mathbf{b}, \quad A\mathbf{x}'' \geqslant \mathbf{b},$$

i.e. $\mathbf{x}' \in R_1$, $\mathbf{x}'' \in R_1$. However,

$$\mathbf{x} = \frac{1}{2}\,\mathbf{x}' + \frac{1}{2}\,\mathbf{x}'',$$

which contradicts the assumption (by the assumption, $\mathbf{x}$ is an extreme point).

(2) Let $(A\mathbf{x})_i = b_i$ for $i = \overline{1,\ r}$ and let $x_j = 0$ for $j = \overline{k+1,\ n}$. Denote

$$\bar{\mathbf{a}}_p = \begin{bmatrix} a_{1p} \\ a_{2p} \\ \ldots \\ a_{rp} \end{bmatrix}, \quad \mathbf{b}' = \begin{bmatrix} b_1 \\ b_2 \\ \ldots \\ b_r \end{bmatrix}, \quad \bar{A} = [\bar{\mathbf{a}}_1,\ \bar{\mathbf{a}}_2,\ \ldots,\ \bar{\mathbf{a}}_k].$$

According to the notation, $\bar{A}\bar{\mathbf{x}} = \mathbf{b}'$ and $\bar{\mathbf{x}} > \mathbf{0}$,

We shall prove that the vectors $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \ldots, \bar{\mathbf{a}}_k$ are linearly independent (in this case $k \leqslant r$).

Assume the contrary, i.e. there is $\bar{\mathbf{x}}' \neq \mathbf{0}$ such that $\bar{A}\bar{\mathbf{x}}' = \mathbf{0}$. We take

$$\mathbf{x}_1 = \begin{bmatrix} \bar{\mathbf{x}} + \varepsilon \bar{\mathbf{x}}' \\ 0 \\ \cdots \\ 0 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} \bar{\mathbf{x}} - \varepsilon \bar{\mathbf{x}}' \\ 0 \\ \cdots \\ 0 \end{bmatrix}.$$

It can easily be seen that $\mathbf{x}_1 \in R_1$ and $\mathbf{x}_2 \in R_1$ for small ε. However,

$$\mathbf{x} = \frac{1}{2}\mathbf{x}_1 + \frac{1}{2}\mathbf{x}_2,$$

which contradicts the assumption. Thus, $k \leqslant r$.

Deleting $r - k$ rows from $\bar{A}$ we obtain a matrix B for which (4.18)-(4.20) hold. $\triangle$

4.3.3. The number of the extreme points of the set R_1 is finite. This assertion is obvious since the number of non-singular square blocks (submatrices) of the condition matrix is finite. $\triangle$

4.3.4. Basis. If $\mathbf{x}^T = (\bar{\mathbf{x}}^T, 0, \ldots, 0)$, $\bar{\mathbf{x}} > \mathbf{0}$, is an extreme point, then the system of linearly independent vectors $\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \ldots, \bar{\mathbf{a}}_k$ in the representation

$$\bar{\mathbf{b}} = \sum_{i=1}^{k} x_i \bar{\mathbf{a}}_i, \quad x_i > 0, \quad i = \overline{1, k},$$

is called the *basis* of the extreme point and the matrix

$$B = [\bar{\mathbf{a}}_1, \bar{\mathbf{a}}_2, \ldots, \bar{\mathbf{a}}_k]$$

is called the *basis matrix* of the extreme point.

4.4. Dual Problems with Mixed Constraints

4.4.1. The most general linear programming problems are problems with mixed constraints where the feasible region is determined by a system of equations and inequalities and part of the variables (or all the variables) can be free of constraints.

We introduce the following notation:

$$I = \{i: i = 1, 2, \ldots, m\}, \quad J = \{j: j = 1, 2, \ldots, n\}.$$

Let I_1 denote part of the set of indices I, i.e. $I_1 \subseteq I$, and let $I_2 = I \setminus I_1$. Similarly, let $J_1 \subseteq J$ and $J_2 = J \setminus J_1$.

Consider a pair of problems

$$\begin{aligned} &\min \langle \mathbf{c}, \mathbf{x} \rangle, \\ &(A\mathbf{x})_i \geqslant b_i, \quad i \in I_1, \\ &(A\mathbf{x})_i = b_i, \quad i \in I_2, \\ &x_j \geqslant 0, \quad j \in J_1, \end{aligned} \tag{4.21}$$

and

$$\begin{aligned} &\max \langle \mathbf{b}, \mathbf{y} \rangle, \\ &(A^T\mathbf{y})_j \leqslant c_j, \quad j \in J_1, \\ &(A^T\mathbf{y})_j = c_j, \quad j \in J_2, \\ &y_i \geqslant 0, \quad i \in I_1. \end{aligned} \tag{4.22}$$

They are called *dual problems with mixed constraints*. As in the case of the problems (4.1) and (4.2), it can readily be shown that the problems (4.21) and (4.22) are mutually dual.

4.4.2. Reduction to equivalent problems of the basic type. We write the problems (4.21) and (4.22) in the following form:

$$\begin{aligned} &\min [\langle \mathbf{c}_1, \mathbf{x}_1 \rangle + \langle \mathbf{c}_2, \mathbf{x}_2 \rangle], \\ &A_{11}\mathbf{x}_1 + A_{12}\mathbf{x}_2 \geqslant \mathbf{b}_1, \\ &A_{21}\mathbf{x}_1 + A_{22}\mathbf{x}_2 = \mathbf{b}_2, \\ &\mathbf{x}_1 \geqslant \mathbf{0}, \end{aligned} \tag{4.23}$$

$$\begin{aligned} &\max [\langle \mathbf{b}_1, \mathbf{y}_1 \rangle + \langle \mathbf{b}_2, \mathbf{y}_2 \rangle], \\ &A_{11}^T\mathbf{y}_1 + A_{21}^T\mathbf{y}_2 \leqslant \mathbf{c}_1, \\ &A_{12}^T\mathbf{y}_1 + A_{22}^T\mathbf{y}_2 = \mathbf{c}_2, \\ &\mathbf{y}_1 \geqslant \mathbf{0}. \end{aligned} \tag{4.24}$$

The notation $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{c}_1$, $\mathbf{c}_2$, $\mathbf{b}_1$, $\mathbf{b}_2$, A_{11}, etc. is understood in a natural way, e.g. $\mathbf{x}_1$ is the vector composed of those components x_j of the vector $\mathbf{x}$ for which $j \in J_1$; to $\mathbf{x}_1$ there corresponds the vector $\mathbf{c}_1$, etc.

The notation of the blocks of the matrix

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is also clear.

In the reduction process we shall replace each equation (e.g. $\varphi = 0$) by two inequalities ($\varphi \geqslant 0$ and $-\varphi \geqslant 0$ or $\varphi \leqslant 0$ and $-\varphi \leqslant 0$), and instead of a variable u on which no nonnegativity conditions are imposed we shall introduce two nonnegative variables $\bar{u}$ and $\bar{\bar{u}}$ in the following way: $u = \bar{u} - \bar{\bar{u}}$ where

$$\bar{u} = \max\{u, 0\} \geqslant 0 \quad \text{and} \quad \bar{\bar{u}} = \max\{-u, 0\} \geqslant 0.$$

Thus, we make the change of variables

$$\mathbf{x}_2 = \bar{\mathbf{x}}_2 - \bar{\bar{\mathbf{x}}}_2, \quad \mathbf{y}_2 = \bar{\mathbf{y}}_2 - \bar{\bar{\mathbf{y}}}_2,$$

$\bar{x}_{j2} = \max\{x_{j2}, 0\}$, $j \in J_2$, $\bar{y}_{i2} = \max\{y_{i2}, 0\}$, $i \in I_2$,
$\bar{\bar{x}}_{j2} = \max\{-x_{j2}, 0\}$, $j \in J_2$, $\bar{\bar{y}}_{i2} = \max\{-y_{i2}, 0\}$, $i \in I_2$.

Introducing these variables and replacing the equations by the corresponding systems of inequalities we reduce (4.23) and (4.24) to

$$\begin{aligned}
&\min[\langle \mathbf{c}_1, \mathbf{x}_1\rangle + \langle \mathbf{c}_2, \bar{\mathbf{x}}_2\rangle - \langle \mathbf{c}_2, \bar{\bar{\mathbf{x}}}_2\rangle], \\
&A_{11}\mathbf{x}_1 + \mathbf{A}_{12}\bar{\mathbf{x}}_2 - \mathbf{A}_{12}\bar{\bar{\mathbf{x}}}_2 \geqslant \mathbf{b}_1, \\
&A_{21}\mathbf{x}_1 + \mathbf{A}_{22}\bar{\mathbf{x}}_2 - \mathbf{A}_{22}\bar{\bar{\mathbf{x}}}_2 \geqslant \mathbf{b}_2, \\
&-A_{21}\mathbf{x}_1 - \mathbf{A}_{22}\bar{\mathbf{x}}_2 + \mathbf{A}_{22}\bar{\bar{\mathbf{x}}}_2 \geqslant -\mathbf{b}_2, \\
&\mathbf{x}_1 \geqslant 0, \quad \bar{\mathbf{x}}_2 \geqslant 0, \quad \bar{\bar{\mathbf{x}}}_2 \geqslant 0,
\end{aligned} \tag{4.25}$$

$$\begin{aligned}
&\max[\langle \mathbf{b}_1, \mathbf{y}_1\rangle + \langle \mathbf{b}_2, \bar{\mathbf{y}}_2\rangle - \langle \mathbf{b}_2, \bar{\bar{\mathbf{y}}}_2\rangle], \\
&A_{11}^T\mathbf{y}_1 + A_{21}^T\bar{\mathbf{y}}_2 - A_{21}^T\bar{\bar{\mathbf{y}}}_2 \leqslant \mathbf{c}_1, \\
&A_{12}^T\mathbf{y}_1 + A_{22}^T\bar{\mathbf{y}}_2 - A_{22}^T\bar{\bar{\mathbf{y}}}_2 \leqslant \mathbf{c}_2, \\
&-A_{12}^T\mathbf{y}_1 - A_{22}^T\bar{\mathbf{y}}_2 + A_{22}^T\bar{\bar{\mathbf{y}}}_2 \leqslant -\mathbf{c}_2, \\
&\mathbf{y}_1 \geqslant 0, \quad \bar{\mathbf{y}}_2 \geqslant 0, \quad \bar{\bar{\mathbf{y}}}_2 \geqslant 0.
\end{aligned} \tag{4.26}$$

It is clear that (4.25) and (4.26) form a pair of dual linear programming problems of the basic type.

4.4.3. It can easily be shown that the problems (4.23) and (4.24) are equivalent to the problems (4.25) and (4.26), namely:

(1) If

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \text{ and } \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

are optimal for (4.23) and (4.24), then

$$\begin{bmatrix} x_1 \\ \overline{x_2} \\ = \\ x_2 \end{bmatrix} \text{ and } \begin{bmatrix} y_1 \\ \overline{y_2} \\ = \\ y_2 \end{bmatrix}$$

are optimal for (4.25) and (4.26), and vice versa.

(2) If the feasible region for (4.23) is empty, then the feasible region for (4.25) is also empty, and vice versa.

(3) If the objective function in the problem (4.23) is unbounded from below on the feasible region, then the objective function in the problem (4.25) is also unbounded from below, and vice versa.

The assertions (1)-(3) are proved with the aid of an elementary argument which, if necessary, the reader can readily carry out.

The equivalence of these problems implies that all the theorems proved for the problems (4.25) and (4.26) remain valid for the problems (4.23) and (4.24) as well.

4.5. Canonical Form of a Linear Programming Problem

4.5.1. The problem

$$\min_{x \in R_0} \langle \mathbf{c}, \mathbf{x} \rangle,$$

$$R_0 = \{\mathbf{x}: A\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geqslant 0\}$$

will be called a *linear programming problem in the canonical form.*

This problem is a special case (where $I_1 = \varnothing$ and $J_2 = \varnothing$) of the problem with mixed constraints (4.23), and, consequently, all the theorems and assertions proved remain valid for the canonical problem as well.

4.5.2. The reduction of the basic problem to the canonical form is performed by introducing additional variables:

for the basic problem

$$\begin{aligned} &\min \langle \mathbf{c}, \mathbf{x} \rangle, \\ &A\mathbf{x} \geqslant \mathbf{b}, \\ &\mathbf{x} \geqslant \mathbf{0} \end{aligned}$$

the equivalent canonical problem is written

$$\begin{aligned} &\min \langle \mathbf{c}, \mathbf{x} \rangle, \\ &A\mathbf{x} - \mathbf{u} = \mathbf{b}, \\ &\mathbf{x} \geqslant \mathbf{0}, \quad \mathbf{u} \geqslant \mathbf{0} \end{aligned}$$

(here the equivalence is understood in the same sense as in problems with mixed constraints).

Noting that the problem dual to the above two problems has the form

$$\begin{aligned} &\max \langle \mathbf{b}, \mathbf{y} \rangle, \\ &A^T\mathbf{y} \leqslant \mathbf{c}, \\ &\mathbf{y} \geqslant \mathbf{0}, \end{aligned}$$

i.e. is the problem (4.2), we see that the proof of the equivalence is trivial.

4.5.3. In what follows we shall always assume that $m < n$ in the canonical problem. This assumption somewhat simplifies the presentation. In this connection we note that in real situations some additional variables forming the so-called *artificial basis* are always introduced, which increases the number of the variables in the problem from n to $n + m$.

4.5.4. **A non-degenerate extreme point.** An extreme point of the set R_0 is said to be *non-degenerate* if its basis matrix is of dimension $m \times m$. Thus, the number of positive components of a non-degenerate extreme point $\mathbf{x}$ is equal to m. For the sake of simplicity, we shall assume that it is the first m components of the extreme point $\mathbf{x}$ that are positive. Then the basis matrix of this non-

degenerate extreme point is

$$B = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{bmatrix}.$$

If, as before, we represent the extreme point x in the form

$$\mathbf{x} = \begin{bmatrix} \bar{\mathbf{x}} \\ 0 \\ \dots \\ 0 \end{bmatrix},$$

where

$$\bar{\mathbf{x}} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_m \end{bmatrix} > 0,$$

then

$$A\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{a}_i = \sum_{i=1}^{m} x_i \mathbf{a}_i = B\bar{\mathbf{x}} = \bar{\mathbf{b}}.$$

4.5.5. Example. Let the set R_0 be determined by

$$\begin{aligned} x_1 + x_2 + x_3 &= 1, \\ x_1 - x_2 &= 0, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0, \quad x_3 &\geqslant 0. \end{aligned}$$

Here R_0 is the line segment joining the points

$$\mathbf{x}_1 = \begin{bmatrix} 1/2 \\ 1/2 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Both $\mathbf{x}_1$ and $\mathbf{x}_2$ are extreme points; however, only one of them, namely $\mathbf{x}_1$, is non-degenerate.

Chapter 5

FINITE METHODS OF SOLVING LINEAR PROGRAMMING PROBLEMS

In this chapter we present the ideas and principal considerations underlying the methods of solving linear programming problems. The reader will not find here a detailed analysis of algorithms. This is due to the following reasons. Knowing the foundations of the method the reader can easily examine a specific algorithm. A detailed presentation of solution algorithms can be found in various books on linear programming. Finally, which is perhaps most important, it is the algorithms related to realization in computer programs that primarily require a thorough analysis. And in this case it is most useful to study special articles and books.

5.1. Simplex Method

5.1.1. The idea of the method. The *simplex method* is a finite iterative scheme consecutively applied to extreme points, under which the value of the objective function decreases from one iteration to another (from one extreme point to another).

We shall consider the canonical linear programming problem

$$\min_{x \in R_0} \langle \mathbf{c}, \mathbf{x} \rangle, \qquad R_0 = \{\mathbf{x}: A\mathbf{x} = \mathbf{b}, \ \mathbf{x} \geqslant \mathbf{0}\}. \tag{5.1}$$

It is assumed that the *problem* (5.1) is *non-degenerate*, i.e. each extreme point of the set R_0 is non-degenerate.

An iteration in the method consists in passing from an extreme point $\mathbf{x}$ to another extreme point $\mathbf{x}'$ so that the value of the objective function decreases:

$$\langle \mathbf{c}, \mathbf{x}' \rangle < \langle \mathbf{c}, \mathbf{x} \rangle.$$

Let an extreme point $\mathbf{x}$ be known. Without loss of generality, we can assume that the basis B of this extreme point is formed of first m columns of the matrix A. We write the matrix A in the following way:

$$A = [B, D],$$

where

$$B = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m], \quad D = [\mathbf{a}_{m+1}, \mathbf{a}_{m+2}, \ldots, \mathbf{a}_n].$$

Accordingly, the vectors $\mathbf{x}$ and $\mathbf{c}$ are written thus:

$$\mathbf{x}^T = (\mathbf{x}_B^T, \mathbf{x}_D^T), \qquad \mathbf{c}^T = (\mathbf{c}_B^T, \mathbf{c}_D^T),$$

$$\mathbf{x}_B^T = (x_1, x_2, \ldots, x_m), \qquad \mathbf{x}_D^T = (x_{m+1}, x_{m+2}, \ldots, x_n),$$

$$\mathbf{c}_B^T = (c_1, c_2, \ldots, c_m), \qquad \mathbf{c}_D^T = (c_{m+1}, c_{m+2}, \ldots, c_n).$$

The components of the vectors $\mathbf{x}_B$ and $\mathbf{x}_D$ are sometimes called *basic* and *nonbasic components*, respectively.

We note that if at least one component of the vector $\mathbf{x}_B$ is zero, then the extreme point $\mathbf{x}$ is degenerate. Indeed, if, for instance, $x_m = 0$, then deleting the last row and column from the matrix B we see that $\mathbf{x}$ satisfies a nonsingular square system of equations of the order $m - 1$, i.e. the extreme point $\mathbf{x}$ is degenerate.

5.1.2. The selection of a column to be introduced in the basis. Let an extreme point $\mathbf{x}$ be known. We have

$$\mathbf{x}_B > \mathbf{0}, \quad \mathbf{x}_D = \mathbf{0}, \quad \det B \neq 0, \quad B\mathbf{x}_B = \mathbf{b}.$$

Consider the vectors

$$\mathbf{x}_k = \mathbf{x}_k(\lambda) = \begin{bmatrix} \mathbf{x}_B - \lambda B^{-1}\mathbf{a}_k \\ 0 \\ \lambda \\ 0 \end{bmatrix}, \quad k = \overline{m+1, n}, \qquad (5.2)$$

where λ is the kth component of the vector $\mathbf{x}$.

We first note that since $\mathbf{x}_B > \mathbf{0}$, we have $\mathbf{x}_k \geqslant \mathbf{0}$ for small $\lambda > 0$. Further, since

$$A\mathbf{x}_k = \sum_{i=1}^{m} [x_i - \lambda (B^{-1}\mathbf{a}_k)_i]\, \mathbf{a}_i + \lambda \mathbf{a}_k$$

$$= B\mathbf{x}_B - \lambda B B^{-1}\mathbf{a}_k + \lambda \mathbf{a}_k = B\mathbf{x}_B = \mathbf{b}, \qquad (5.3)$$

$\mathbf{x}_k \in R_0$ for small $\lambda > 0$. Moreover,

$$\langle \mathbf{c}, \mathbf{x}_k \rangle = \langle \mathbf{c}_B, \mathbf{x}_B \rangle - \lambda \langle \mathbf{c}_B, B^{-1}\mathbf{a}_k \rangle + \lambda c_k$$
$$= \langle \mathbf{c}, \mathbf{x} \rangle - \lambda [\langle \mathbf{c}_B, B^{-1}\mathbf{a}_k \rangle - c_k].$$

Denote

$$\Delta_k = \langle \mathbf{c}_B, B^{-1}\mathbf{a}_k \rangle - c_k. \tag{5.4}$$

The parameter Δ_k is defined for any $k = \overline{1, n}$ and for $k = \overline{1, m}$ we have

$$\Delta_k = \langle \mathbf{c}_B, B^{-1}\mathbf{a}_k \rangle - c_k = \langle \mathbf{c}_B, \mathbf{e}_k \rangle - c_k = c_k - c_k = 0$$

since $B^{-1}\mathbf{a}_k = \mathbf{e}_k$ ($\mathbf{e}_k$ is the kth coordinate vector). Finally, we obtain

$$\langle \mathbf{c}, \mathbf{x}_k \rangle = \langle \mathbf{c}, \mathbf{x} \rangle - \lambda \Delta_k, \quad k = \overline{1, n}. \tag{5.5}$$

5.1.3. The selection of a column to be removed from the basis. Depending on the signs of Δ_k and $(B^{-1}\mathbf{a}_k)_i$, there arise the following three cases.

I. If $\Delta_k \leqslant 0$ for any $k = \overline{1, n}$, then the point $\mathbf{x}$ is optimal.

Indeed, denote

$$\bar{\mathbf{y}} = (B^{-1})^T \mathbf{c}_B.$$

Then the condition $\Delta_k \leqslant 0$ is written

$$\Delta_k = \langle \mathbf{c}_B, B^{-1}\mathbf{a}_k \rangle - c_k = \langle (B^{-1})^T \mathbf{c}_B, \mathbf{a}_k \rangle - c_k$$
$$= \langle \bar{\mathbf{y}}, \mathbf{a}_k \rangle - c_k \leqslant 0, \quad k = \overline{1, n},$$

or, which is the same,

$$A^T \bar{\mathbf{y}} \leqslant \mathbf{c}.$$

Since the problem dual to (5.1) is

$$\max_{\mathbf{y} \in Q_0} \langle \mathbf{b}, \mathbf{y} \rangle, \quad Q_0 = \{\mathbf{y}: A^T\mathbf{y} \leqslant \mathbf{c}\}, \tag{5.6}$$

we have $\bar{\mathbf{y}} \in Q_0$. And since

$$\langle \mathbf{c}, \mathbf{x} \rangle = \langle \mathbf{c}_B, \mathbf{x}_B \rangle = \langle \mathbf{c}_B, B^{-1}\mathbf{b} \rangle = \langle (B^{-1})^T \mathbf{c}_B, \mathbf{b} \rangle = \langle \bar{\mathbf{y}}, \mathbf{b} \rangle,$$

Theorem 4.2.5 implies the optimality of the point $\mathbf{x}$.

II. If there is an index $k \geqslant m + 1$ such that

$$\Delta_k > 0 \quad \text{and} \quad B^{-1}\mathbf{a}_k \leqslant \mathbf{0},$$

then the set R_0 is unbounded and the function $\langle \mathbf{c}, \mathbf{x} \rangle$ is unbounded from below on R_0.

Indeed, it follows from (5.2) that for any $\lambda > 0$ we have

$$\mathbf{x}_k = \mathbf{x}_k(\lambda) \geqslant \mathbf{0}.$$

This and (5.3) imply that $\mathbf{x}_k(\lambda) \in R_0$. Therefore (5.3) implies that

$$\langle \mathbf{c}, \mathbf{x}_k(\lambda) \rangle \to -\infty$$

for $\lambda \to +\infty$.

III. Let there be $k \geqslant m+1$ and $i \leqslant m$ such that

$$\Delta_k > 0 \quad \text{and} \quad (B^{-1}\mathbf{a}_k)_i > 0^*.$$

In this case an iteration is performed. Denote

$$I_k = \{i: (B^{-1}\mathbf{a}_k)_i > 0\}$$

and take

$$\lambda = \lambda_0 = \min_{i \in I_k} \frac{(B^{-1}\mathbf{b})_i}{(B^{-1}\mathbf{a}_k)_i} = \frac{(B^{-1}\mathbf{b})_s}{(B^{-1}\mathbf{a}_k)_s}. \tag{5.7}$$

We note that $\lambda_0 > 0$ since

$$(B^{-1}\mathbf{b})_i = x_i > 0, \quad i = \overline{1, m}.$$

For this choice of λ_0 the point $\mathbf{x}_k = \mathbf{x}_k(\lambda_0)$ becomes a feasible point of the set R_0 (see (5.2) and (5.3)). The index s for which the minimum is attained in (5.7) is unique due to the non-degeneracy of the problem (5.1). Indeed, if the minimum were attained for more than one index, (5.3) would imply that the vector $\mathbf{b}$ is expressed as a linear combination with positive coefficients of less than m vectors $\mathbf{a}_i$ of the matrix A, i.e. the point $\mathbf{x}_k$ would satisfy a system of equations of dimension less than m, and, consequently, $\mathbf{x}_k$ would be a degenerate extreme point. However, this contradicts the assumption of non-degeneracy. The fact that $\mathbf{x}_k$ is an extreme point can easily be verified. To this end it suffices to show that the system of vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m, \mathbf{a}_k$ is linearly independent. Assume the contrary, i.e. let there be some numbers $\beta_1, \beta_2, \ldots, \beta_{s-1}, \beta_{s+1}, \ldots,$

---

* We remind the reader that $(B^{-1}\mathbf{a}_k)_i$ denotes the ith component of the vector $B^{-1}\mathbf{a}_k$.

β_m, β_k not all zero such that

$$\sum_{\substack{i=1 \\ i \neq s}}^{m} \beta_i \mathbf{a}_i + \beta_k \mathbf{a}_k = 0. \tag{5.8}$$

Denote

$$z_{ik} = (B^{-1}\mathbf{a}_k)_i. \tag{5.9}$$

Then the identity $BB^{-1}\mathbf{a}_k - \mathbf{a}_k \equiv 0$ is written thus:

$$\sum_{i=1}^{m} z_{ik}\mathbf{a}_i - \mathbf{a}_k \equiv 0.$$

Multiplying the identity by β_k and adding the result to (5.8) we find

$$\sum_{\substack{i=1 \\ i \neq s}}^{m} (\beta_i + \beta_k z_{ik})\,\mathbf{a}_i + \beta_k z_{sk}\mathbf{a}_s = 0.$$

Since $\det B \neq 0$, the above relation holds if and only if all the coefficients are zero:

$$\beta_i + \beta_k z_{ik} = 0, \quad i = \overline{1,\ m},\ i \neq s, \quad \beta_k z_{sk} = 0.$$

However,

$$z_{sk} = (B^{-1}\mathbf{a}_k)_s > 0,$$

since $s \in I_k$, and, consequently, $\beta_k = 0$, and therefore all β_i are zero, which contradicts the assumption that the numbers $\beta_1, \beta_2, \ldots, \beta_{s-1}, \beta_{s+1}, \ldots, \beta_m, \beta_k$ are not all zero.

5.1.4. The finiteness of the method. Thus, let $\mathbf{x}_k$ be a new extreme point such that

$$\langle \mathbf{c},\ \mathbf{x}_k \rangle = \langle \mathbf{c},\ \mathbf{x} \rangle - \lambda_0 \Delta_k < \langle \mathbf{c},\ \mathbf{x} \rangle. \tag{5.10}$$

What was presented above implies that an iteration in the simplex method consists in passing from a basis $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_s, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m$ to the basis $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m, \mathbf{a}_k$ so that the objective function decreases. Hence, the simplex method results in an extreme point $\mathbf{x}^*$ where the minimum is attained after a finite number of iterations.

5.2. Recurrence Relations of the Simplex Method Algorithm (Relationship between the Parameters of Successive Iterations)

5.2.1. We introduce the following notation:

$$\mathbf{a}_0 = \mathbf{b}, \quad z_{ik} = (B^{-1}\mathbf{a}_k)_i, \quad i = \overline{1,\ m}, \quad k = \overline{0,\ n},$$

$$z_{0j} = \Delta_j, \quad j = \overline{1,\ n}, \quad z_{00} = \langle \mathbf{c},\ \mathbf{x} \rangle.$$

Thus, the parameters corresponding to an extreme point $\mathbf{x}$ are the numbers z_{ij} $(i = \overline{0,\ m};\ j = \overline{0,\ n})$ and the parameters of the new extreme point $\mathbf{v} \underline{\underline{\triangle}} \mathbf{x}_k\ (\lambda_0)$ are the numbers v_{ij}. The relationship between the parameters is expressed by the following formulas:

$$\begin{aligned} v_{ij} &= z_{ij} - z_{ik}\frac{z_{sj}}{z_{sk}}, \quad i = \overline{0,\ m}, \quad i \neq s, \quad j = \overline{0,\ n}, \\ v_{hj} &= \frac{z_{sj}}{z_{sk}}, \qquad\qquad\qquad\quad j = \overline{0,\ n}. \end{aligned} \tag{5.11}$$

Indeed,

$$\mathbf{a}_k = B\,(B^{-1}\mathbf{a}_k) = \sum_{i=1}^{m} z_{ik}\mathbf{a}_i,$$

and, since, according to the condition $z_{sk} = (B^{-1}\mathbf{a}_k)_s > 0$, we have

$$\mathbf{a}_s = \frac{1}{z_{sk}}\,\mathbf{a}_k - \sum_i{}' \frac{z_{ik}}{z_{sk}}\,\mathbf{a}_i,$$

where

$$\sum_i{}' \underline{\underline{\triangle}} \sum_{\substack{i=1 \\ i \neq s}}^{m}.$$

Further, since

$$\mathbf{a}_j = \sum_{i=1}^{m} z_{ij}\mathbf{a}_i, \quad j = \overline{0,\ n},$$

we have

$$\mathbf{a}_j = \sum_i{}' z_{ij}\mathbf{a}_i + z_{sj}\mathbf{a}_s =.$$

$$= \sum_i{}' z_{ij}\mathbf{a}_i + z_{sj}\left(\frac{1}{z_{sk}}\,\mathbf{a}_k - \sum_i{}' \frac{z_{ik}}{z_{sk}}\,\mathbf{a}_i\right)$$

$$= \sum_i{}' \left(z_{ij} - z_{ik}\frac{z_{sj}}{z_{sk}}\right)\mathbf{a}_i + \frac{z_{sj}}{z_{sk}}\,\mathbf{a}_k.$$

Comparing this representation of the vector $\mathbf{a}_j$ with the relations

$$\mathbf{a}_j = \sum_i{}' v_{ij}\mathbf{a}_i + v_{kj}\mathbf{a}_k,$$

determining the parameters v_{ij} we obtain the formulas (5.11) for $j = \overline{0, n}$ and $i = \overline{1, m}$. For $i = 0$ and $j = \overline{1, n}$ we have

$$z_{0j} = \langle \mathbf{c}_B,\ B^{-1}\mathbf{a}_j\rangle - c_j = \sum_{i=1}^{m} c_i z_{ij} - c_j.$$

Therefore

$$v_{0j} = \sum_i{}' c_i v_{ij} + c_k v_{kj} - c_j$$

$$= \sum_i{}' c_i \left(z_{ij} - z_{ik}\frac{z_{sj}}{z_{sk}}\right) - c_j + c_k\frac{z_{sj}}{z_{sk}}$$

$$= \sum_{i=1}^{m} c_i \left(z_{ij} - z_{ik}\frac{z_{sj}}{z_{sk}}\right) - c_j + c_k\frac{z_{sj}}{z_{sk}}$$

$$= z_{0j} - \left(\sum_{i=1}^{m} c_i z_{ik} - c_k\right)\frac{z_{sj}}{z_{sk}} = z_{0j} - z_{0k}\frac{z_{sj}}{z_{sk}}.$$

Finally, (5.10) and (5.7) imply

$$v_{00} = z_{00} - \lambda_0\Delta_k = z_{00} - \frac{z_{s0}}{z_{sk}}\,z_{0k}.$$

Thus, the validity of (5.11) is proved.

5.3. Methods of Finding an Initial Extreme Point

5.3.1. The most widespread methods are the so-called *artificial basis methods* whose essence lies in the following: an artificial auxiliary canonical linear programming problem with an extreme point known in advance is constructed whose solution is used to find in a trivial manner

either an extreme point or an optimal point for the problem (5.1). Thus, the process of finding an initial extreme point is related to the solution of a new linear programming problem, and therefore, generally, it is as labour-consuming as the process of finding the solution of the problem (5.1) with a given initial extreme point.

5.3.2. The artificial basis method. The problem of finding an extreme point of the set

$$R_0 = \{\mathbf{x}: A\mathbf{x} = \mathbf{b},\ \mathbf{x} \geqslant \mathbf{0}\}$$

is considered. Without loss of generality, we can assume that $\mathbf{b} \geqslant \mathbf{0}$. We construct the following auxiliary problem in the space E_{n+m}:

$$\min_{\mathbf{x},\mathbf{u}\in W} \sum_{i=1}^{m} u_i, \qquad W = \left\{ \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} : A\mathbf{x} + \mathbf{u} = \mathbf{b},\ \mathbf{x} \geqslant \mathbf{0},\ \mathbf{u} \geqslant \mathbf{0} \right\} \tag{5.12}$$

for which the extreme point

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

of the set W is known in advance. Applying the simplex method we find the solution $\begin{bmatrix} \mathbf{x}^* \\ \mathbf{u}^* \end{bmatrix}$ to the problem (5.12). Denote

$$\mu = \sum_{i=1}^{m} u_i^*.$$

5.3.3. Theorem. *If* $\mu = 0$, *then* $\mathbf{x}^*$ *is an extreme point of the set* R_0. *If* $\mu > 0$, *then* $R_0 = \varnothing$.

Proof. First, it is clear that the problem (5.12) is solvable since $W \neq \varnothing$ and the objective function $\sum_{i=1}^{m} u_i$ is bounded from below by zero.

If $\mu = 0$, then

$$\begin{bmatrix} \mathbf{x}^* \\ \mathbf{u}^* \end{bmatrix} \quad \begin{bmatrix} \mathbf{x}^* \\ \mathbf{0} \end{bmatrix}$$

is an optimal extreme point of the problem (5.12) since this problem is solved by the simplex method in which

successive iterations are performed for extreme points. It is obvious that $\mathbf{x}^*$ is an extreme point of the set R_0. Now let $\mu > 0$. Assume that $R_0 \neq \varnothing$, i.e. there is $\mathbf{x} \in R_0$. Then

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{0} \end{bmatrix}$$

is the solution to the problem (5.12), which contradicts the assumption that μ is positive. $\triangle$

We note that the artificial basis method is included as a subprogram in the majority of standard computer programs for the simplex method of solving linear programming problems.

In the application of the artificial basis method the problem (5.1) has to be solved in two stages: first, the auxiliary problem (5.12) is solved, and then the problem (5.1) itself is dealt with. The method below makes it possible to combine the two stages.

5.3.4. **The M-method.** Consider the following auxiliary problem:

$$\begin{gathered} \min_{\mathbf{x},\mathbf{u}\in W} \left[\langle \mathbf{c}, \mathbf{x} \rangle + M \sum_{i=1}^{m} u_i \right], \\ W = \left\{ \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} : A\mathbf{x} + \mathbf{u} = \mathbf{b}, \quad \mathbf{x} \geqslant \mathbf{0}, \quad \mathbf{u} \geqslant \mathbf{0} \right\}. \end{gathered} \tag{5.13}$$

5.3.5. **Theorem.** *If the problem* (5.1) *is solvable, then there is a number M_0 such that for all $M > M_0$ in any solution $\begin{bmatrix} \mathbf{x}^* \\ \mathbf{u}^* \end{bmatrix}$ of the problem* (5.13) *the point $\mathbf{x}^*$ is optimal for the problem* (5.1).

Proof. According to the statement of the theorem, the problem (5.1) is solvable, i.e. there is an optimal point $\mathbf{x}^*$. Then the dual problem

$$\langle \mathbf{b}, \mathbf{y}^* \rangle = \max_{\mathbf{y}\in Q_0} \langle \mathbf{b}, \mathbf{y} \rangle,$$

$$Q_0 = \{\mathbf{y}: A^T\mathbf{y} \leqslant \mathbf{c}\}$$

is also solvable. By Theorem 4.2.5, $\mathbf{x}^*$ and $\mathbf{y}^*$ must satisfy the following system:

$$A\mathbf{x} = \mathbf{b}, \quad \mathbf{x} \geqslant \mathbf{0}, \quad A^T\mathbf{y} \leqslant \mathbf{c}, \quad \langle \mathbf{c}, \mathbf{x} \rangle = \langle \mathbf{b}, \mathbf{y} \rangle. \tag{5.14}$$

Since the problem

$$\max_{y \in P} \langle \mathbf{b}, \mathbf{y} \rangle, \quad P = \{\mathbf{y}: A^T \mathbf{y} \leqslant \mathbf{c}, \ y_i \leqslant M, \ i = \overline{1, m}\}$$

is dual to (5.13), the necessary and sufficient condition for existence of the solution to the problem (5.13) (and, consequently, to its dual problem) is the existence of the solution to the following system of inequalities:

$$\begin{gathered} A\mathbf{x} + \mathbf{u} = \mathbf{b}, \quad \mathbf{x} \geqslant \mathbf{0}, \quad \mathbf{u} \geqslant \mathbf{0}, \\ A^T \mathbf{y} \leqslant \mathbf{c}, \quad y_i \leqslant M, \quad i = \overline{1, m}, \\ \langle \mathbf{c}, \mathbf{x} \rangle + M \sum_{i=1}^{m} u_i = \langle \mathbf{b}, \mathbf{y} \rangle. \end{gathered} \tag{5.15}$$

Since $\mathbf{x}^*$ and $\mathbf{y}^*$ satisfy the system (5.14), we see that $\mathbf{x} = \mathbf{x}^*$, $\mathbf{u} = \mathbf{u}^* = \mathbf{0}$, and $\mathbf{y} = \mathbf{y}^*$ satisfy the system (5.15) for all

$$M > M_0 = \max\{y_1^*, y_2^*, \ldots, y_m^*\}.$$

The fact that for $M > M_0$ we have $\mathbf{u}^* = \mathbf{0}$ in any solution $\mathbf{x}^*$, $\mathbf{u}^*$ to the problem (5.13) and, consequently, $\mathbf{x}^*$ is the solution to the problem (5.1) is a consequence of the conditions (4.6). △

5.3.6. Discussion. It is obvious that for $\mathbf{b} \geqslant \mathbf{0}$ the point

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{b} \end{bmatrix}$$

is an extreme point of the set W. It is this point that is chosen as the initial point in the application of the simplex method.

In practice there is no need in determining the number M_0. Usually a sufficiently large number is taken as M, e.g.

$$M > \max_{\substack{i=\overline{1, n} \\ j=\overline{1, m}}} \{|a_{ij}|, \ |c_i|, \ |b_j|\}.$$

Finally, we note that if the M-problem (5.13) possesses a solution, then the solution of the system of inequalities (5.15) exists, and, consequently, $Q_0 \neq \varnothing$. Therefore, Theorem 4.2.8 implies: if $R_0 \neq \varnothing$ and there exists a

solution to the M-problem, then the original problem (5.1) is solvable. Consequently, Theorem 5.3.5 is true under the conditions that the solution to the M-problem exists and the set R_0 in nonempty. We thus arrive at the following conclusion: if there are arbitrarily large values of M such that in the solution $\begin{bmatrix} \mathbf{x}^* \\ \mathbf{u}^* \end{bmatrix}$ to the problem (5.13) we have

$$\max_{i=\overline{1, m}} u_i^* > 0,$$

then the original problem (5.1) is unsolvable.

5.4. Degeneracy. Perturbation Method

5.4.1. Cycling. In the case when an extreme point $\mathbf{x}$ of the set R_0 is degenerate, the number of its positive components is less than m. Then it may turn out that $\lambda_0 = 0$ (see (5.7)), and, consequently, when passing in the simplex method from the degenerate extreme point $\mathbf{x}$ to the subsequent extreme point, the value of the objective function may not decrease (see (5.10)). Moreover, as was seen, the degeneracy may result in ambiguity in the selection of the vector removed from the basis when passing from the extreme point $\mathbf{x}$ to a degenerate extreme point $\mathbf{x}_k = \mathbf{x}_k(\lambda_0)$.

Theoretically, these factors may lead to cycling when after a number of iterations the procedure of the simplex method returns to the previous extreme point. To overcome the cycling phenomenon, there is a rule for selecting the vector removed from the basis with the aid of the so-called perturbation method.

5.4.2. The perturbation method. Assume that to an extreme point $\mathbf{x}$ of the set R_0 there corresponds a linearly independent system of m vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ such that $\mathbf{x}_B = B^{-1}\mathbf{b} \geqslant \mathbf{0}$ (in contrast to the non-degenerate case when $\mathbf{x}_B > \mathbf{0}$), where $B = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$.

Consider the "perturbed" vector

$$\mathbf{b}(\varepsilon) = \mathbf{b} + \sum_{j=1}^{n} \varepsilon^j \mathbf{a}_j,$$

where ε is a sufficiently small number ($\varepsilon > 0$) and ε_j is the jth power of ε. Denoting, as before, $z_{ij} = (B^{-1}\mathbf{a}_j)_i$

we find

$$\mathbf{b}+\sum_{j=1}^{n}\varepsilon^j\mathbf{a}_j=\sum_{i=1}^{m}x_i\mathbf{a}_i+\sum_{j=1}^{n}\varepsilon^j\sum_{i=1}^{m}z_{ij}\mathbf{a}_i$$

$$=\sum_{i=1}^{m}\left(x_i+\sum_{j=1}^{n}z_{ij}\varepsilon^j\right)\mathbf{a}_i.$$

Since

$$z_{ij}=\begin{cases} z_{ij}, & j=\overline{m+1,\ n},\\ 0, & j\neq i,\ j=\overline{1,\ m},\\ 1, & j=i,\ j=\overline{1,\ m},\end{cases}$$

we have

$$x_i(\varepsilon)\underset{=}{\triangle}x_i+\sum_{j=1}^{n}z_{ij}\varepsilon^j=x_i+\varepsilon^i+\sum_{j=m+1}^{n}z_{ij}\varepsilon^j,\quad i=\overline{1,\ m}. \tag{5.16}$$

5.4.3. The ε-problem. Along with the problem

$$\min_{\mathbf{x}\in R_0}\langle\mathbf{c},\ \mathbf{x}\rangle,$$
$$R_0=\{\mathbf{x}:\ A\mathbf{x}=\mathbf{b},\ \mathbf{x}\geqslant 0\} \tag{5.1}$$

we shall consider the so-called ε-problem

$$\min_{\mathbf{x}(\varepsilon)\in R_0(\varepsilon)}\langle\mathbf{c},\ \mathbf{x}(\varepsilon)\rangle,$$
$$R_0(\varepsilon)=\{\mathbf{x}(\varepsilon):\ A\mathbf{x}(\varepsilon)=\mathbf{b}(\varepsilon),\ \mathbf{x}(\varepsilon)\geqslant 0\} \tag{5.17}$$

and show that this problem is non-degenerate for all sufficiently small $\varepsilon>0$.

Let $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$ be an arbitrary linearly independent system of vectors of the matrix A. Since the functions $x_i(\varepsilon)$ $(i=\overline{1,\ m})$ are not identically zero polynomials in ε (see (5.16), each of them has no less than n roots, and, consequently, no more than n positive roots.

Let η be the least positive root among all the positive roots of the m polynomials $x_i(\varepsilon)$ $(i=\overline{1,\ m})$ (if none of the polynomials possesses positive roots, we put $\eta=\infty$). The root η is determined by the chosen system of linearly

independent vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m$, and, since the number of linearly independent systems of vectors of the matrix A is finite, the number of the corresponding least positive roots η is also finite.

Let $\varepsilon_0 > 0$ be the smallest of all the numbers η (i.e. the least positive root of all the polynomials $x_i(\varepsilon)$ constructed from all the possible linearly independent systems of vectors of the matrix A). It is obvious that for $\varepsilon < \varepsilon_0$ none of the polynomials (5.16) vanishes ($\varepsilon > 0$), and, consequently, each extreme point of the ε-problem contains m positive components, i.e. is non-degenerate.

It can be proved that, for small ε, to the extreme points $\mathbf{x} \in R_0$ and $\mathbf{x}(\varepsilon) \in R_0(\varepsilon)$ there corresponds one and the same basis, and therefore, to choose in a unique manner an index $s \in I_k$ for which the minimum of the ratio $(B^{-1}\mathbf{b})_i/(B^{-1}\mathbf{a}_k)_i$ is attained (see (5.7)) in the simplex procedure of solving the problem (5.1), it is advisable to turn to the ε-problem. Thus, as the index s of the vector $\mathbf{a}_s$ to be removed from the basis one should take the index (which is unique due to the non-degeneracy of the ε-problem) for which the minimum of the ratio

$$\frac{x_s(\varepsilon)}{(B^{-1}\mathbf{a}_k)_s} = \min_{i \in I_k} \frac{(B^{-1}\mathbf{b}(\varepsilon))_i}{(B^{-1}\mathbf{a}_k)_i}$$

$$= \min_{i \in I_k} \left[\frac{x_j}{z_{ik}} + \varepsilon \frac{z_{i1}}{z_{ik}} + \varepsilon^2 \frac{z_{i2}}{z_{ik}} + \ldots + \varepsilon^n \frac{z_{in}}{z_{ik}} \right]$$

is attained.

5.4.4. In practice there is no need in passing to the ε-problem. The above formula clearly shows that the choice of the index s can be carried out in the following way.

Denote by $I_{k0}(s_k)$ the set of the indices $s_k \in I_k$ for which the minimum of the ratio

$$\frac{x_{s_k}}{z_{s_k k}} = \min_{i \in I_k} \frac{x_i}{z_{ik}}, \quad s_k \in I_{k0}(s_k)$$

is attained. If $I_{k0}(s_k)$ consists of more than one element, we calculate

$$\frac{z_{s_l 1}}{z_{s_l k}} = \min_{s_k \in I_{k0}(s_k)} \frac{z_{s_k 1}}{z_{s_k k}}, \quad s_l \in I_{k1}(s_l).$$

If $I_{k1}(s_l)$ consists of a single element, then s_l is the sought-for index. If $I_{k1}(s_l)$ contains two or more elements, we calculate

$$\frac{z_{s_p 2}}{z_{s_p k}} = \min_{s_l \in I_{k1}(s_l)} \frac{z_{s_l 2}}{z_{s_l k}},$$

etc. It is obvious that this process is finite and leads to a unique choice of the desired index.

5.4.5. The experience in the application of the simplex method has shown that to eliminate the possibility of cycling it suffices to use any regular technique for choosing the index s in a unique manner, e.g. as s one can take the first (in the order of the chosen indexing) index for which the minimum of the ratio (5.7) is attained. And it seems natural that the index k must also be chosen (if it is not unique) in a regular way; e.g. among all k such that $\Delta_k > 0$ and $(B^{-1}\mathbf{a}_k)_i > 0$ we can take the one for which Δ_k is minimum.

5.5. On the Simplex Method for Solving Special Classes of Linear Programming Problems

5.5.1. Since any linear programming problem, whatever the form it is written in, can be reduced to an equivalent canonical problem, the simplex method is, in a sense, a universal method in linear programming.

However, as a rule, the reduction to the canonical form brings about a strong increase of the dimension of the condition matrix. And this, first, requires a great computer internal storage capacity (which is substantially limited) and, second, strongly increases the number of iterations to obtain the solution to the problem.

All this has made it necessary to separate out classes of frequently encountered typical problems and to elaborate special versions of the simplex method for them.

These classes of problems first of all include problems with two-sided constraints on the variables when the feasible region is determined by conditions of the form $A\mathbf{x} = \mathbf{b}$, $\mathbf{d}_1 \leqslant \mathbf{x} \leqslant \mathbf{d}_2$.

5.6. On the Modified Simplex Method

5.6.1. In the ordinary simplex method all the elements of the matrix A are transformed at each iteration according to the formulas (5.11). In the modified method the matrix B^{-1} is calculated at each iterative step and then, using B^{-1}, the parameters

$$\bar{\mathbf{y}} = (B^{-1})^T \mathbf{c}_B, \quad \Delta_k = \langle \bar{\mathbf{y}}, \mathbf{a}_k \rangle - c_k, \quad B^{-1}\mathbf{b}$$

are calculated, etc. After this, with the aid of simple recurrence relations one passes from the matrix

$$B^{-1} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_s, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m]^{-1}$$

to the matrix

$$\bar{B}^{-1} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_k, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m]^{-1}$$

of the corresponding new extreme point.

Thus, at a given iteration only the elements of the matrix B^{-1} are transformed and stored in the computer memory. Here the amount of current information memorized by the computer is, as a rule, reduced. Moreover, when the matrix A contains a great number of zero elements, the scheme in the modified simplex method can be compiled so that in the calculation of scalar products at each step only nonzero elements of A take part. This substantially reduces the amount of calculations and allows one to place the nonzero elements of the matrix A in the computer storage compactly.

5.6.2. The recurrence relations of the modified simplex method can be derived from the formulas for transformation of the matrix B^{-1} into the matrix $\bar{B}^{-1}$.

Denote the elements of the matrices B^{-1} and $\bar{B}^{-1}$ by a_{ij}^{-1} and $\bar{a}_{ij}^{-1}$, respectively:

$$B^{-1} = [a_{ij}^{-1}], \quad \bar{B}^{-1} = [\bar{a}_{ij}^{-1}].$$

Then

$$\bar{a}_{ij}^{-1} = a_{ij}^{-1} - z_{ik} \frac{a_{sj}^{-1}}{z_{sk}}, \quad i \neq s,$$

$$\bar{a}_{sj}^{-1} = \frac{a_{sj}^{-1}}{z_{sk}}.$$

The validity of these formulas is verified by multiplying the matrix $\bar{B}$ by $\bar{B}^{-1}$, which results in the unit matrix.

5.7. Multiplicative Representation of an Inverse Matrix

5.7.1. Multipliers. To reduce the amount of information to be stored, in the modified simplex method use is often made of the following multiplicative representation of an inverse matrix.

By an *elementary matrix* we shall mean a matrix of the form

$$G_s = \begin{bmatrix} 1 & 0 & \dots & w_{1k} & 0 & \dots & 0 \\ 0 & 1 & \dots & w_{2k} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{sk} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & w_{mk} & 0 & \dots & 1 \end{bmatrix},$$

where

$$w_{ik} = -\frac{z_{ik}}{z_{sk}}, \quad i \neq s, \quad w_{sk} = \frac{1}{z_{sk}}$$

(the index s shows that the column of the elements w_{jk} occupies the sth place). It can easily be verified that

$$G_s B^{-1} = \bar{B}^{-1}.$$

The procedure of the modified simplex method begins, as a rule, with a unit basis to which a unit matrix G corresponds. At the first step let a vector $\mathbf{a}_{k_1}$ be introduced in the basis to replace $\mathbf{e}_{s_1}$. Then the inverse matrix B_1^{-1} of the matrix of the new basis is

$$B_1^{-1} = G_{s_1} G,$$

and at the pth step

$$B_p^{-1} = G_{s_p} G_{s_{p-1}} \dots G_{s_1} G$$

Since each matrix G_s is determined by $(m + 1)$ parameters s, w_{1k}, w_{2k}, . . ., w_{mk}, the multiplicative form of the inverse matrix occupies a smaller part of the computer storage than the ordinary form.

5.8. Dual Simplex Method*

5.8.1. The idea of the method. As was seen, the simplex method allows one to obtain, along with the solution to the primal problem, the solution

$$\bar{\mathbf{y}} = (B^{-1})^T \mathbf{c}_B$$

to the dual problem as well (case I of the simplex method). It is this result that underlies the dual simplex method of solving the problem (5.1).

The essence of the method consists in a process of successive iterations for extreme points of the feasible region Q_0 of the dual problem (5.6) such that the value of the objective function $\langle \mathbf{b}, \mathbf{y} \rangle$ increases from one extreme point to another, i.e. in the application of the simplex method to solving the dual problem. Since we are interested in the dual simplex method as a method of solving (5.1), we shall elucidate the meaning in (5.1) of the transformation performed at each iterative step of the simplex method applied to the problem (5.6).

We shall assume that the problem (5.6) is non-degenerate, i.e. to each extreme point of the set Q_0 there corresponds a square non-degenerate system of equations of dimension m. The matrix of the system will be called the *dual basis of the primal problem* (5.1).

5.8.2. Iterative procedure of the method. Let an extreme point $\mathbf{y}$ of the set Q_0 be known. This means that we know a dual basis

$$B = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m]$$

such that the point $\mathbf{y} = (B^{-1})^T \mathbf{c}_B$ satisfies the inequality

$$\Delta = A^T \mathbf{y} - \mathbf{c} \leqslant \mathbf{0}$$

and the relations

$$\Delta_i = (\mathbf{y}, \mathbf{a}_i) - c_i = (A^T \mathbf{y})_i - c_i = \begin{cases} 0 & \text{for } i = \overline{1, m}, \\ \Delta_i < 0 & \text{for } i = \overline{m+1, n} \end{cases}$$

hold.

* Along with the name "dual simplex method", the term "method of consecutive refinement of estimates" is used. Its meaning will be elucidated in Sec. 5.9.

As before, denote

$$\mathbf{x}_B = B^{-1}\mathbf{b}, \quad z_{ik} = (B^{-1}\mathbf{a}_k)_i$$

and, moreover,

$$B^{-1} \underline{\underline{\triangle}} [\mathbf{a}_1^{-1}, \ \mathbf{a}_2^{-1}, \ \ldots, \ \mathbf{a}_m^{-1}],$$

i.e. $\mathbf{a}_i^{-1}$ designates the ith row of the matrix B^{-1}.

Consider the vectors

$$\mathbf{y}_s \underline{\underline{\triangle}} \mathbf{y}_s(\theta) = \mathbf{y} - \theta \mathbf{a}_s^{-1}, \quad s = \overline{1, \ m}.$$

Since

$$\langle \mathbf{a}_s^{-1}, \ \mathbf{a}_i \rangle = \begin{cases} 0 & \text{for} \quad i = \overline{1, \ m}, \quad i \neq s, \\ 1 & \text{for} \quad i = s, \\ z_{si} & \text{for} \quad i = \overline{m+1, \ n}, \end{cases}$$

we have

$$(A^T \mathbf{y}_s)_i = \langle \mathbf{y}_s, \ \mathbf{a}_i \rangle$$

$$= \begin{cases} c_i & \text{for} \quad i = \overline{1, \ m}, \quad i \neq s, \\ c_s - \theta & \text{for} \quad i = s, \\ \langle \mathbf{y}, \ \mathbf{a}_i \rangle - \theta z_{si} = \Delta_i - \theta z_{si} + c_i & \text{for} \quad i = \overline{m+1, \ n}. \end{cases} \tag{5.18}$$

Depending on the signs of x_j and z_{ij}, there arise the following three cases.

I. If $\mathbf{x}_B \geqslant \mathbf{0}$, then

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_B \\ 0 \\ \ldots \\ 0 \end{bmatrix} \in R_0,$$

and, consequently, the point $\mathbf{x}$ is optimal (see case I in the simplex method).

II. If there is an index $s \leqslant m$ such that $x_s < 0$ and $z_{si} \geqslant 0$ for all $i = \overline{m+1, \ n}$, then $\mathbf{y}_s \in Q_0$ for any $\theta > 0$, and, consequently, the set Q_0 is unbounded and

$$\langle \mathbf{b}, \ \mathbf{y}_s \rangle = \langle \mathbf{b}, \ \mathbf{y} \rangle - \theta \langle \mathbf{b}, \ \mathbf{a}_s^{-1} \rangle = \langle \mathbf{b}, \ \mathbf{y} \rangle - \theta x_s \to +\infty$$

for $\theta \to +\infty$, i.e. the set R_0 is empty (see 4.2.9).

III. Let there be $s \leqslant m$ and $i \geqslant m+1$ such that $x_s < 0$ and $z_{si} < 0$. In this case an iterative step is performed. Denote $I_s = \{i:\ z_{si} < 0\}$ and take

$$\theta = \theta_0 = \min_{i \in I_s} \frac{\Delta_i}{z_{si}} > 0.$$

Let

$$\theta_0 = \frac{\Delta_k}{z_{sk}}, \quad k \in I_s. \tag{5.19}$$

The index k for which this minimum is attained is unique due to the non-degeneracy of the problem (5.6)*.

Let us prove that

$$\mathbf{y}_s = \mathbf{y}_s(\theta_0) = \mathbf{y} - \theta_0 \mathbf{a}_s^{-1} \in Q_0.$$

Indeed, (5.18) and (5.19) imply

$$(A^T \mathbf{y}_s)_i \begin{cases} = c_i & \text{for } i = \overline{1, m}, \quad i \neq s, \\ < c_s & \text{for } i = s, \\ < c_i & \text{for } i = \overline{m+1, n}, \quad i \neq k, \\ = c_k & \text{for } i = k. \end{cases}$$

Finally,

$$\langle \mathbf{b}, \mathbf{y}_s \rangle = \langle \mathbf{b}, \mathbf{y} \rangle - \theta_0 \langle \mathbf{b}, \mathbf{a}_s^{-1} \rangle = \langle \mathbf{b}, \mathbf{y} \rangle - \theta_0 x_s > \langle \mathbf{b}, \mathbf{y} \rangle.$$

Thus, when passing from the point $\mathbf{y}$ to the point $\mathbf{y}_s$ the value of the objective function of the problem (5.6) increases.

5.8.3. Discussion. We see that an iterative step in the dual simplex method consists in that a vector $\mathbf{a}_s$ is removed from the dual basis and $\mathbf{a}_k$ is introduced in place of it; under the iteration the value of the objective function increases. It remains to prove that the new system of vectors $\mathbf{a}_1, \ldots, \mathbf{a}_{s-1}, \mathbf{a}_k, \mathbf{a}_{s+1}, \ldots, \mathbf{a}_m$ is linearly independent, i.e. forms a new dual basis. The reader can readily show this in a way similar to that in case III of the simplex method.

Since the number of the extreme points of the set Q_0 is finite and the value of the objective function increases at each iteration (when passing from one extreme point to

* The proof of this assertion is similar to the proof of an analogous assertion in case III of the simplex method.

another), after a finite number of iterations the method leads to the solution $\mathbf{y}^*$ of the dual problem and, consequently, to the solution $\mathbf{x}^*$ of the primal problem.

The recurrence relations of the dual simplex method algorithm can easily be derived from the formulas for transition from the matrix B^{-1} to the matrix $\bar{B}^{-1}$ (see Sec. 5.6).

In the case of a degenerate problem, just like in the ordinary simplex method, use is made of certain rules elaborated for the dual simplex method and based on the perturbation method, which make it possible to uniquely select the index k for θ_0, i.e. rules for selecting a new vector to be introduced in the dual basis.

5.9. Solution of the Dual Problem Interpreted as Estimation of Influence

5.9.1. We shall show that under certain variations of the components of the constraint vector $\mathbf{b}$ in the problem (5.1) the components of the solution to the dual problem (5.6) can be interpreted as estimates for the influence of these variations on the minimum of the objective function in (5.1).

Consider the problem

$$\begin{gathered}\min_{\mathbf{x}\in R_0(\mathbf{z})} \langle \mathbf{c}, \mathbf{x}\rangle, \\ R_0(\mathbf{z}) = \{\mathbf{x}: A\mathbf{x} = \mathbf{z}, \quad \mathbf{x} \geqslant \mathbf{0}\}\end{gathered} \tag{5.20}$$

and its dual problem

$$\begin{gathered}\max_{\mathbf{y}\in Q_0} \langle \mathbf{z}, \mathbf{y}\rangle, \\ Q_0 = \{\mathbf{y}: A^T\mathbf{y} \leqslant \mathbf{c}\}.\end{gathered} \tag{5.21}$$

Assuming that the problem (5.20) is solvable for some $\mathbf{z}$ we denote

$$\mu(\mathbf{z}) = \min_{\mathbf{x}\in R_0(\mathbf{z})} \langle \mathbf{c}, \mathbf{x}\rangle.$$

Let for $\mathbf{z} = \mathbf{z}^*$ there exist a solution $\mathbf{x}^*$ to the problem (5.20) and, consequently, there exist a solution $\mathbf{y}^*$ to the problem (5.21).

5.9.2. Theorem. *If the problem* (5.20) *is non-degenerate, then there is a neighborhood* $U(\mathbf{z}^*)$ *of the point* $\mathbf{z}^*$ *such that*

$$\mu(\mathbf{z}) = \langle \mathbf{z}, \mathbf{y}^* \rangle$$

for all $\mathbf{z} \in U(\mathbf{z}^*)$.

Proof. Since the problem (5.20) is solvable for $\mathbf{z} = \mathbf{z}^*$, we see that, according to Theorem 4.2.12, there is an optimal extreme point $\mathbf{x}^*$. Let $\mathbf{x}^*$ have the form

$$\mathbf{x}^* = \begin{vmatrix} \mathbf{x}_B^* \\ 0 \\ \dots \\ 0 \end{vmatrix}$$

where

$$\mathbf{x}_B^* = \begin{bmatrix} x_1^* \\ x_2^* \\ \dots \\ x_m^* \end{bmatrix} > \mathbf{0}.$$

To $\mathbf{x}^*$ there corresponds a basis $B = [\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m]$, $\det B \neq 0$, and $\mathbf{x}_B^* = B^{-1}\mathbf{z}^*$. Denote

$$\alpha = \min_{i=\overline{1, m}} x_i^*.$$

We take the following neighborhood:

$$U(\mathbf{z}^*) = \left\{ \mathbf{z}: \| \mathbf{z} - \mathbf{z}^* \| \leqslant \frac{\alpha}{\| B^{-1} \|} \right\}.$$

We shall show that the point

$$\mathbf{x}^T = (\mathbf{x}_B^T, 0, \dots, 0),$$

where $\mathbf{x}_B = B^{-1}\mathbf{z}$, is optimal for the problem (5.20) for any $\mathbf{z} \in U(\mathbf{z}^*)$. Indeed,

$$\| \mathbf{x}_B^* - \mathbf{x}_B \| = \| B^{-1}(\mathbf{z}^* - \mathbf{z}) \| \leqslant \| B^{-1} \| \, \| \mathbf{z}^* - \mathbf{z} \| \leqslant \alpha,$$

where the norm is defined thus:

$$\| \mathbf{x}_B^* - \mathbf{x}_B \| = \max_{i=\overline{1, m}} |x_i^* - x_i|.$$

We see that

$$x_i^* - x_i \leqslant \alpha, \quad i = \overline{1, m},$$

i.e.

$$x_j \geqslant x_j^* - \alpha = x_j^* - \min_{i=\overline{1,\ m}} x_i^* \geqslant 0, \quad j = \overline{1,\ m}.$$

Thus, $\mathbf{x} \geqslant \mathbf{0}$. Moreover, by construction, $A\mathbf{x} = \mathbf{z}$, i.e. $\mathbf{x} \in R_0(\mathbf{z})$.

We shall prove that $\mathbf{x}$ is optimal for the problem (5.20). It follows from the conditions (4.6) that to each $x_i^* > 0$ $(i = \overline{1,\ m})$ there corresponds

$$(A^T\mathbf{y}^*)_i = c_i.$$

Therefore,

$$\langle \mathbf{c},\ \mathbf{x}\rangle = \sum_{i=1}^{m} c_i x_i = \sum_{i=1}^{m} (A^T\mathbf{y}^*)_i\, x_i = \sum_{i=1}^{m} \langle \mathbf{a}_i,\ \mathbf{y}^*\rangle\, x_i$$

$$= \left\langle \mathbf{y}^*,\ \sum_{i=1}^{m} \mathbf{a}_i x_i \right\rangle = \langle \mathbf{y}^*,\ B\mathbf{x}_B\rangle = \langle \mathbf{y}^*,\ \mathbf{z}\rangle.$$

Now, since $\mathbf{x} \in R_0(\mathbf{z})$ and $\mathbf{y}^* \in Q_0$, Theorem 4.2.5 implies that the points $\mathbf{x}$ and $\mathbf{y}^*$ are optimal and

$$\mu(\mathbf{z}) = \langle \mathbf{y}^*,\ \mathbf{z}\rangle. \quad \triangle$$

5.9.3. It is obvious that

$$\frac{\partial \mu(\mathbf{z})}{\partial z_i} = y_i^*,$$

and therefore an increase of z_i results in an increase or decrease of $\mu(\mathbf{z})$ depending on the sign of y_i^*. Here the rate of change of $\mu(\mathbf{z})$ is determined by the parameter $|y_i^*|$.

The dual simplex method is also called the *method of consecutive refinement of estimates* since the extreme points in (5.6) appearing in the iterations can be regarded as approximate values of the exact estimate of $\mathbf{y}^*$, i.e. as approximate estimates of the influence of the conditions of the problem (5.1) on the minimum of the objective function.

5.10. On the Dual Simplex Method for Problems with Increasing Number of Constraints

5.10.1. Let the optimal extreme point $\mathbf{x}^*$ and $\mathbf{y}^*$ of the respective problems (5.1) and (5.6) be found. Assume that to the system of conditions determining the feasible region

R_0 one more inequality is added:

$$\sum_{i=1}^{n} \alpha_i x_i \geqslant b_{m+1}.$$

If

$$\sum_{i=1}^{n} \alpha_i x_i^* \geqslant b_{m+1},$$

then $\mathbf{x}^*$ will be optimal for the new problem as well since the addition of a constraint can only increase the minimum of the objective function. In the case when

$$\sum_{i=1}^{n} \alpha_i x_i^* < b_{m+1}$$

we shall find out how the solution of the new problem can be obtained from the known optimal extreme points $\mathbf{x}^*$ and $\mathbf{y}^*$.

We write the new problem in the canonical form by introducing an additional variable x_{n+1}. Denote

$$\boldsymbol{\alpha}^T = (\alpha_1, \alpha_2, \ldots, \alpha_n), \quad \bar{A} = \begin{bmatrix} A \\ \boldsymbol{\alpha}^T \end{bmatrix}, \quad \bar{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ x_{n+1} \end{bmatrix},$$

$$\bar{\mathbf{c}} = \begin{bmatrix} \mathbf{c} \\ 0 \end{bmatrix}, \quad \bar{\mathbf{b}} = \begin{bmatrix} \mathbf{b} \\ b_{m+1} \end{bmatrix}, \quad \bar{\mathbf{y}} = \begin{bmatrix} \mathbf{y} \\ y_{m+1} \end{bmatrix},$$

and let $\mathbf{e}_{m+1}$ be the $(m+1)$th unit coordinate vector. The new problem takes the form

$$\begin{gathered} \min \langle \bar{\mathbf{c}}, \bar{\mathbf{x}} \rangle \\ [\bar{A}, -\mathbf{e}_{m+1}] \bar{\mathbf{x}} = \bar{\mathbf{b}}, \quad \bar{\mathbf{x}} \geqslant 0. \end{gathered} \tag{5.22}$$

According to the rules (4.23)-(4.24) the problem dual to (5.22) is

$$\begin{gathered} \max [\langle \mathbf{b}, \mathbf{y} \rangle + b_{m+1} y_{m+1}], \\ A^T \mathbf{y} + \boldsymbol{\alpha} y_{m+1} \leqslant \mathbf{c}, \\ y_{m+1} \geqslant 0. \end{gathered} \tag{5.23}$$

Since $\mathbf{y}^*$ is an extreme point for the problem (5.6),

$$\bar{\mathbf{y}}^* = \begin{bmatrix} \mathbf{y}^* \\ 0 \end{bmatrix}$$

is an extreme point for the problem (5.23) and the linear independence of the system of vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$ (the basis of the extreme point $\mathbf{y}^*$) implies that the system of vectors

$$\begin{vmatrix} \mathbf{a}_1 \\ \alpha_1 \end{vmatrix}, \begin{vmatrix} \mathbf{a}_2 \\ \alpha_2 \end{vmatrix}, \ldots, \begin{vmatrix} \mathbf{a}_m \\ \alpha_m \end{vmatrix}, \begin{vmatrix} \mathbf{0} \\ 1 \end{vmatrix}$$

is also linearly independent. Now it is natural to solve the problem (5.22) by the dual simplex method taking $\mathbf{y}^*$ as an initial extreme point. As a rule, to find the solution to (5.22) a small number of iterations is needed.

It is obvious that all that has been said in this section remains true when to the inequalities determining the set R_0 not one but several new inequalities are added.

5.11. Method of Simultaneous Solution of the Primal and Dual Problems

5.11.1. The methods considered above involved two stages. At the first stage it was required to find an extreme point of the feasible region, at the second stage the procedure of determining an optimal point was constructed using that extreme point. Generally, both stages are equally labor-consuming. The iterative process of the method of simultaneous solution of the primal and dual problems begins with an arbitrary feasible (not necessarily extreme) point of the dual problem whose determination, in real situations, can often be performed quite easily. For instance, if $\mathbf{c} \geqslant \mathbf{0}$, then $\mathbf{y} = \mathbf{0}$ belongs to the set Q_0.

Thus, let us consider the canonical primal problem (5.1)

$$\min_{\mathbf{x} \in R_0} \langle \mathbf{c}, \mathbf{x} \rangle$$

$$R_0 = \left\{ \mathbf{x}: \sum_{i=1}^{n} x_i \mathbf{a}_i = \mathbf{b}, \ \mathbf{x} \geqslant \mathbf{0}, \ \mathbf{b} \geqslant \mathbf{0} \right\}$$

and its dual problem (5.6):

$$\max_{\mathbf{y} \in Q_0} \langle \mathbf{b}, \mathbf{y} \rangle,$$

$$Q_0 = \{ \mathbf{y}: \langle \mathbf{a}_j, \mathbf{y} \rangle \leqslant c_j, \ j = \overline{1, n} \}.$$

Assume that a point $\mathbf{y}_0 \in Q_0$ is known. We denote

$$J = \{ j: \ j = \overline{1, n} \}, \quad J(\mathbf{y}_0) = \{ j: \ \langle \mathbf{a}_j, \mathbf{y}_0 \rangle = c_j, \ j \in J \}$$

and consider the so-called *extended problem*

$$\min \sum_{i=1}^{m} \varepsilon_i,$$

$$\sum_{j=1}^{n} x_j \mathbf{a}_j + \varepsilon = \mathbf{b}, \tag{5.24}$$

$$\mathbf{x} \geqslant \mathbf{0}, \quad \varepsilon \geqslant \mathbf{0},$$

where $\varepsilon^T = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m)$, and "truncate" it in the following way:

$$\min \sum_{i=1}^{m} \varepsilon_i,$$

$$\sum_{j \in J(\mathbf{y}_0)} z_j \mathbf{a}_j + \varepsilon = \mathbf{b}, \tag{5.25}$$

$$z_j \geqslant 0, \quad j \in J(\mathbf{y}_0), \quad \varepsilon \geqslant 0.$$

The theorem below lies in the foundation of the method under consideration.

5.11.2. Theorem. *If* $\varepsilon^* = 0$ *in the solution* $\mathbf{z}^*$, ε^* *to the problem* (5.25), *then the point* $\mathbf{x}^*$ *such that*

$$x_j^* = \begin{cases} z_j^*, & j \in J(\mathbf{y}_0), \\ 0, & j \in J \setminus J(\mathbf{y}_0) \end{cases}$$

is optimal for the problem (5.1) *and the point* $\mathbf{y}_0$ *is optimal for the problem* (5.6).

Proof. It is obvious that $\mathbf{x}^* \in R_0$. Since $\mathbf{y}_0$ is a feasible point for (5.6) and $\varepsilon^* = 0$, we find

$$\langle \mathbf{c}, \mathbf{x}^* \rangle = \sum_{j \in J(\mathbf{y}_0)} c_j x_j^* + \sum_{j \in J \setminus J(\mathbf{y}_0)} c_j x_j^* = \sum_{j \in J(\mathbf{y}_0)} \mathbf{c}_j z_j^*$$

$$= \sum_{j \in J(\mathbf{y}_0)} \langle \mathbf{a}_j, \mathbf{y}_0 \rangle z_j^* = \left\langle \mathbf{y}_0, \sum_{j \in J(\mathbf{y}_0)} z_j^* \mathbf{a}_j \right\rangle = \langle \mathbf{y}_0, \mathbf{b} \rangle.$$

Since $\mathbf{x}^* \in R_0$, $\mathbf{y}_0 \in Q_0$, and $\langle \mathbf{c}, \mathbf{x}^* \rangle = \langle \mathbf{b}, \mathbf{y}_0 \rangle$, Theorem 4.2.5 implies that $\mathbf{x}^*$ and $\mathbf{y}_0$ are optimal. △

5.11.3. The iterative process in the method. We construct the problem dual to (5.25):

$$\max \langle \mathbf{b}, \mathbf{u} \rangle$$

$$\langle \mathbf{a}_j, \mathbf{u} \rangle \leqslant 0, \quad j \in J(\mathbf{y}_0), \tag{5.26}$$

$$u_i \leqslant 1, \quad i = \overline{1, m}.$$

The problems (5.25) and (5.26) are solvable. Indeed, since $\mathbf{b} \geqslant \mathbf{0}$, we see that $z_j = 0$, $j \in J(\mathbf{y}_0)$, and $\varepsilon = \mathbf{b}$ are feasible for (5.25), $\mathbf{u} = \mathbf{0}$ is a feasible point for (5.26), and, by Theorem 4.2.8, both problems are solvable.

Let $\mathbf{u}^*$ be the solution to (5.26). Denote

$$\Delta^0 = A^T\mathbf{y}_0 - \mathbf{c}, \quad \delta^* = A^T\mathbf{u}^*.$$

The vectors Δ^0 and δ^* are often called *discrepancy vectors*.

Depending on the properties of the parameters ε_i^* ($i = \overline{1, m}$) and δ_j^* ($j \in J$), there arise the following three cases.

I. If

$$\sum_{i=1}^{m} \varepsilon_i^* = 0,$$

then, by Theorem 5.11.2, the point $\mathbf{x}^*$ is a solution to (5.1).

II. Let

$$\sum_{i=1}^{m} \varepsilon_i^* > 0$$

and let there be an index $j \in J \setminus J(\mathbf{y}_0)$ for which $\delta_j^* > 0$. In this case we construct a new point $\mathbf{y} \in Q_0$ to which a greater value of the objective function of the problem (5.6) corresponds. We take

$$\mathbf{y} \underline{\underline{\Delta}} \mathbf{y}(\mu) = \mathbf{y}_0 - \mu\mathbf{u}^*.$$

Then

$$\Delta \underline{\underline{\Delta}} A^T\mathbf{y} - \mathbf{c} = A^T\mathbf{y}_0 - \mu A^T\mathbf{u}^* - \mathbf{c} = \Delta^0 - \mu\delta^*. \tag{5.27}$$

We note that

$$\Delta_j^0 = \begin{cases} 0, & j \in J(\mathbf{y}_0), \\ \Delta_j^0 < 0, & j \in J \setminus J(\mathbf{y}_0), \end{cases}$$
$$\delta_j^* = \begin{cases} \delta_j^* \leqslant 0, & j \in J(\mathbf{y}_0), \\ \delta_j^*, & j \in J \setminus J(\mathbf{y}_0). \end{cases} \tag{5.28}$$

Taking

$$\mu_0 = \frac{\Delta_k^0}{\delta_k^*} = \max_{\delta_j^* > 0} \frac{\Delta_j^0}{\delta_j^*} < 0 \tag{5.29}$$

we see that

$$\Delta^0 - \mu_0 \delta^* \leqslant 0,$$

and, consequently, by virtue of (5.27), $\mathbf{y} = \mathbf{y}(\mu_0) \in Q_0$. Since $\mathbf{z}^*$, $\boldsymbol{\varepsilon}^*$ is the solution to the problem (5.25) and $\mathbf{u}^*$ is the solution to the dual problem of (5.26), we have

$$\langle \mathbf{b}, \mathbf{u}^* \rangle = \sum_{i=1}^{m} \varepsilon_i^*$$

(by Theorem 4.2.3). Therefore

$$\langle \mathbf{b}, \mathbf{y} \rangle = \langle \mathbf{b}, \mathbf{y}_0 \rangle - \mu_0 \langle \mathbf{b}, \mathbf{u}^* \rangle = \langle \mathbf{b}, \mathbf{y}_0 \rangle - \mu_0 \sum_{i=1}^{m} \varepsilon_i^* > \langle \mathbf{b}, \mathbf{y}_0 \rangle$$

since $\mu_0 < 0$ and $\sum_{i=1}^{m} \varepsilon_i^* > 0$.

Thus, an iterative step in the method consists in solving the truncated problem (5.25), calculating the parameter μ_0, and constructing a new point $\mathbf{y}_0 = \mathbf{y}_0(\mu_0)$. In the first iteration the point with the components $z_j = 0$, $j \in J(\mathbf{y}_0)$, and $\varepsilon_i = b_i \geqslant 0$ $(i = \overline{1, m})$ is taken as the initial extreme point of the problem (5.25), after which the simplex method is applied.

5.11.4. In the next iterative step a new truncated problem is constructed:

$$\begin{gathered} \min \sum_{i=1}^{m} \varepsilon_i, \\ \sum_{j \in J(\mathbf{y})} z_j \mathbf{a}_j + \boldsymbol{\varepsilon} = \mathbf{b}, \\ z_j \geqslant 0, \quad j \in J(\mathbf{y}), \quad \boldsymbol{\varepsilon} \geqslant \mathbf{0}, \end{gathered} \tag{5.30}$$

and then the process is repeated. As the initial extreme point of the problem (5.30) one takes a point whose non-zero coordinates are the positive components of the optimal extreme point $\mathbf{z}^*$, $\boldsymbol{\varepsilon}^*$ of the problem (5.25) since

$$\{j: z_j^* > 0\} \subseteq J(\mathbf{y}).$$

This inclusion relation follows from the fact that to the values $z_j^* > 0$ there correspond

$$\delta_j^* = \langle \mathbf{a}_j, \mathbf{u}^* \rangle = 0$$

(this is implied by relations of the type (4.6) in application to the solutions of (5.25) and (5.26), and, consequently (see (5.27) and (5.28)),

$$\Delta_j = \langle \mathbf{a}_j, \mathbf{y} \rangle - c_j = \Delta_j^0 - \mu_0 \delta_j^* = 0.$$

However, by construction,

$$J(\mathbf{y}) = \{j: \Delta_j = \langle \mathbf{a}_j, \mathbf{y} \rangle - c_j = 0\},$$

which is what we intended to prove.

The procedure of solving the problem (5.30) begins with the introduction of a vector $\mathbf{a}_k$ in the basis (the index k is taken from (5.29)). We note that $k \in J(\mathbf{y})$ since $\Delta_k = 0$ (see (5.29)). As a result, the sum $\sum_{i=1}^{m} \varepsilon_i$ becomes smaller than $\sum_{i=1}^{m} \varepsilon_i^*$. Indeed, the parameters

$$x_j^* = \begin{cases} z_j^*, & j \in J(\mathbf{y}_0), \\ 0, & j \in J \setminus J(\mathbf{y}_0), \end{cases} \quad \text{and} \quad \varepsilon_i^*, \ i = \overline{1, m}$$

determine an extreme point of the feasible region of the problem (5.24). To the introduction of $\mathbf{a}_k$ in the basis of (5.30) there corresponds an ordinary iterative step of the simplex method when solving (5.24) (see (5.29)). However, when the problem (5.24) is non-degenerate, the objective function decreases at each iteration in the simplex method, i.e. $\sum_{i=1}^{m} \varepsilon_i$ becomes smaller than $\sum_{i=1}^{m} \varepsilon_i^*$. It follows that at each step of the method of simultaneous solution of the primal and dual problems the optimal value of the objective function $\sum_{i=1}^{m} \varepsilon_i$ decreases as compared to its optimal value in the foregoing truncated problem. This implies that the method is finite. We also note that since for each problem of the type (5.30) there is an initial extreme point and the index set $J(\mathbf{y})$ changes insignificantly from step to step, the number of iterations in the simplex method of solving the problem (5.30) is not great.

III. If $\sum_{i=1}^{m} \varepsilon_i^* > 0$ and $\delta^* \leqslant 0$, then $\mathbf{y} = \mathbf{y}(\mu) = \mathbf{y}_0 - \mu \mathbf{u}^* \in Q_0$ for any $\mu < 0$ since

$$\Delta = A^T \mathbf{y} - \mathbf{c} = \Delta^0 - \mu \delta^* \leqslant \Delta^0 = A^T \mathbf{y}_0 - \mathbf{c} \leqslant 0.$$

However,

$$\langle \mathbf{b}, \mathbf{y} \rangle = \langle \mathbf{b}, \mathbf{y}_0 \rangle - \mu \langle \mathbf{b}, \mathbf{u}^* \rangle = \langle \mathbf{b}, \mathbf{y}_0 \rangle - \mu \sum_{i=1}^{m} \varepsilon_i^* \to \infty$$

for $\mu \to -\infty$, i.e. $\langle \mathbf{b}, \mathbf{y} \rangle$ is unbounded on Q_0, and, consequently, $R_0 = \varnothing$ (see 4.2.9).

Thus, the method of simultaneous solution of the primal and dual problems consists in constructing a sequence of solutions to truncated problems until case I arises and the solution is found or case III is realized, which shows that $R_0 = \varnothing$.

5.12. Decomposition Method

5.12.1. As a rule, practical linear programming problems contain a large number of the unknowns and constraints. From the viewpoint of computer realization of the solution methods for such problems it seems advantageous to reduce the solution of a problem of a great complexity to the solution of several problems of lesser extent. The decomposition method is a representative of this group of methods.

5.12.2. A block problem. Notation:

Vectors:

$$\mathbf{c}_j^T = (c_{j1}, c_{j2}, \ldots, c_{jn_j}),$$
$$\mathbf{x}_j^T = (x_{j1}, x_{j2}, \ldots, x_{jn_j}),$$
$$\mathbf{b}_0^T = (b_{01}, b_{02}, \ldots, b_{0m}),$$
$$\mathbf{b}_j^T = (b_{j1}, b_{j2}, \ldots, b_{jm_j}).$$

Matrices:

$$A_j \text{ (dimension } m \times n_j),$$
$$B_j \text{ (dimension } m_j \times n_j).$$

Here and henceforth $j = 1, 2, \ldots, n$.

By a *block problem* is meant the following: find

$$\min \sum_{j=1}^{n} \langle \mathbf{c}_j, \mathbf{x}_j \rangle \tag{5.31}$$

under the conditions

$$\sum_{j=1}^{n} A_j \mathbf{x}_j = \mathbf{b}_0, \tag{5.32}$$

$$B_j \mathbf{x}_j = \mathbf{b}_j, \tag{5.33}$$

$$\mathbf{x}_j \geqslant \mathbf{0}. \tag{5.34}$$

5.12.3. The extremal problem. Consider the following sets:

$$S_j = \{\mathbf{x}_j: B_j \mathbf{x}_j = \mathbf{b}_j, \ \mathbf{x}_j \geqslant \mathbf{0}\}. \tag{5.35}$$

We first assume that all the sets S_j are bounded.

Denote by W_j the set of the extreme points $\mathbf{x}_j^h$ of the convex polyhedron S_j:

$$W_j = \{\mathbf{x}_j^1, \ \mathbf{x}_j^2, \ \ldots, \ \mathbf{x}_j^{h_j}\}.$$

We also denote

$$\mathbf{p}_j^h = A_j \mathbf{x}_j^h, \quad \mathbf{c}_j^h = \langle \mathbf{c}_j, \mathbf{x}_j^h \rangle, \quad h = \overline{1, h_j}. \tag{5.36}$$

The problem (5.31)-(5.34) is now reduced to the so-called *extremal problem on extreme points*. According to Theorem 2.2.10, any point $\mathbf{x}_j \in S_j$ can be represented as a convex combination of extreme points:

$$\mathbf{x}_j = \sum_{h=1}^{h_j} \mathbf{x}_j^h v_{jh}, \quad \sum_{h=1}^{h_j} v_{jh} = 1, \quad v_{jh} \geqslant 0, \quad h = \overline{1, h_j}. \tag{5.37}$$

As a result of the substitution of $\mathbf{x}_j$ from (5.37) into (5.31)-(5.34) and by virtue of the notation (5.36), there arises the following *extremal problem*: find

$$\min \sum_{j=1}^{n} \sum_{h=1}^{h_j} \mathbf{c}_j^h v_{jh} \tag{5.38}$$

under the conditions

$$\sum_{j=1}^{n} \sum_{h=1}^{h_j} \mathbf{p}_j^h v_{jh} = \mathbf{b}_0, \tag{5.39}$$

$$\sum_{k=1}^{k_j} v_{jk} = 1, \quad j = \overline{1, n}, \tag{5.40}$$

$$v_{jk} \geqslant 0, \quad j = \overline{1, n}, \quad k = 1, k_j. \tag{5.41}$$

It is obvious that if v_{jk} is a solution to the problem (5.38)-(5.41), then the vectors $\mathbf{x}_j = \sum_{k=1}^{k_j} \mathbf{x}_j^k v_{jk}$ are a solution to the problem (5.31)-(5.34) (remember the notation (5.36) and the condition $B_j \mathbf{x}_j^k = \mathbf{b}_j$ holding since $\mathbf{x}_j^k \in S_j$).

We note that the block problem involves $m + \sum_{j=1}^{n} m_j$ equations and $\sum_{i=1}^{n} n_j$ unknowns. In the extremal problem the number of equations is reduced to $m + n$ while the number of unknowns substantially increases and becomes equal to $\sum_{j=1}^{n} k_j$, i.e. to the total number of the extreme points of all the polyhedrons S_j.

5.12.4. The algorithm. The algorithm for the extremal problem makes it possible to use, in the active stage of the solution, only a small part of total number $\sum_{j=1}^{n} k_j$ of the parameters without calculating in advance all the elements in (5.38)-(5.41).

We shall solve the problem (5.38)-(5.41) by the modified simplex method. Let the matrix of the initial basis P of the extremal problem consisting of $m + n$ vectors having the form $\begin{bmatrix} \mathbf{p}_i^l \\ \mathbf{e}_i \end{bmatrix}$, the parameters corresponding to this basis in the conditions (5.39)-(5.40), and the vectors

$$(\mathbf{y}^T, \bar{\mathbf{y}}^T) = (P^{-1})^T \bar{\mathbf{c}},$$

where $\bar{\mathbf{c}}^T = (\ldots c_{il} \ldots)$, be known.

An ordinary iterative step in the simplex method for solving (5.38)-(5.41) consists in finding the indices (j, k) for which the inequality $\Delta_{jk} = \langle \mathbf{y}, \mathbf{p}_j^k \rangle + \langle \bar{\mathbf{y}}, \mathbf{e}_j \rangle - c_j^k = \bar{y}_j + \langle \mathbf{y}, \mathbf{p}_j^k \rangle - c_j^k > 0$ holds, introducing a vector

$\begin{bmatrix} \mathbf{p}_j^k \\ \mathbf{e}_j \end{bmatrix}$ into the basis P, and removing a vector from P so that the new system forms a basis. The process ends when all $\Delta_{jk} \leqslant 0$. However, to determine the signs of the parameters Δ_{jk} by the standard method it would be necessary to test a great number of these parameters $\left(\text{which is equal to } \sum_{j=1}^{n} k_j - (m+n)\right)$, and all $\mathbf{p}_j^k$ and hence all the extreme points $\mathbf{x}_j^k$ must be known.

The algorithm provides a procedure of selecting Δ_{jk} such that at each step one must solve n linear programming problems of a small extent. Following this rule, one finds the greatest value Δ^* among all possible Δ_{jk}, and if $\Delta^* \leqslant 0$, the process is stopped and, as a result, the optimal solution is determined. In case $\Delta^* > 0$, the next calculation step is carried out.

In order to find Δ^* one should find in each polyhedron S_j an extreme point $\mathbf{x}_j^{r_j}$ such that

$$\Delta_{jr_j} = \max_{W_j} \Delta_{jk},$$

and then take the greatest value among all Δ_{jr_j}, i.e. Δ^*.

From the construction of the sets S_j and W_j and the fact that

$$\Delta_{jk} = \bar{y}_j + \langle \mathbf{y},\ \mathbf{p}_j^k \rangle - c_j^k = \bar{y}_j + \langle A_j^T \mathbf{y} - \mathbf{c}_j,\ \mathbf{x}_j^k \rangle$$

it follows that

$$\Delta_{jr_j} = \max_{W_j} \Delta_{jk} = \max_{S_j} (\bar{y}_j + \langle A_j^T \mathbf{y} - \mathbf{c}_j,\ \mathbf{x}_j \rangle).$$

Thus, an iterative step consists in the following.

For each index j an *auxiliary problem* of finding

$$\max \langle A_j^T \mathbf{y} - \mathbf{c}_j,\ \mathbf{x}_j \rangle \tag{5.42}$$

under the conditions

$$B_j \mathbf{x}_j = \mathbf{b}_j, \quad \mathbf{x}_j \geqslant \mathbf{0}$$

is solved. Denote by $\mathbf{x}_j^{r_j}$ the solution to this problem. We thus obtain a sequence of solutions $\mathbf{x}_1^{r_1}, \mathbf{x}_2^{r_2}, \ldots, \mathbf{x}_n^{r_n}$.

After that all $\Delta_{jr_j}(\mathbf{x}_j^{r_j})$ $(j = \overline{1, n})$ are computed, and the index i is found for which

$$\max_{j=\overline{1,n}} \Delta_{jr_j}(\mathbf{x}_j^{r_j}) = \Delta_{ir_i}(\mathbf{x}_i^{r_i}) = \Delta^*$$

is attained.

If $\Delta^* \leqslant 0$, the construction of the solution ends.

If $\Delta^* > 0$, then the vector $\begin{bmatrix} A_i \mathbf{x}_i^{r_i} \\ \mathbf{c}_i \end{bmatrix}$ is introduced into the basis P of the extremal problem and a column vector is removed from it according to the rules of the modified simplex method.

5.12.5. The beginning of the algorithm. To find the matrix of the initial basis for the problem (5.38)-(5.41), the artificial basis method is applied (see 5.3.2). One introduces $m + n$ artificial variables $u_i \geqslant 0$ $(i = \overline{1, m+n})$ and solves the following problem: find

$$\min \sum_{i=1}^{m+n} u_i$$

under the conditions

$$\begin{gathered} \sum_{j=1}^{n} \sum_{k=1}^{k_j} \mathbf{p}_j^k v_{jk} + \sum_{i=1}^{m} \mathbf{c}_i u_i = \mathbf{b}_0, \\ \sum_{k=1}^{k_j} v_{jk} + u_{m+j} = 1, \quad j = \overline{1, n}, \\ v_{jk} \geqslant 0, \quad u_l \geqslant 0, \quad j = \overline{1, n}; \quad k = \overline{1, k_j}; \quad l = \overline{1, m+n}. \end{gathered} \tag{5.43}$$

For $\mathbf{b}_0 \geqslant \mathbf{0}$ (this does not lead to loss of generality) the matrix of the initial basis is a unit matrix of dimension $m \times n$ to which the values $(u_1, u_2, \ldots, u_{m+n}) = (\mathbf{b}_0^T, 1, 1, \ldots, 1)$ of the basic variables correspond.

Further, the problem is solved with the aid of the procedure described in the foregoing section.

5.12.6. Extension of the method to the case of an unbounded set S_j. As was seen, under the assumption that the sets S_j are bounded the presented algorithm leads to two possible situations, namely, $\Delta^* \leqslant 0$, when the solution to the problem is found, or $\Delta^* > 0$, when the next iteration is performed.

If the assumption that the sets S_j are bounded is dropped, then there can be a case when at the current step

of the simplex method the objective function is unbounded from above on an unbounded feasible region (see case II in Sec. 5.1.3).

To derive relations indicating the unboundedness from above of the objective function of (5.42) one should remember that an iteration in the simplex method (in the notation of Secs. 5.1.2 and 5.1.3) consists in passing from an already known extreme point $\bar{\mathbf{x}}$ with a basis B to a new point $\mathbf{x}$ with the aid of the formula $\mathbf{x} = \bar{\mathbf{x}} - \lambda \mathbf{z}$, where

$$\mathbf{z} = \begin{bmatrix} -B^{-1}\mathbf{a}_h - \\ 0 \\ -1 \\ \cdots \\ 0 \end{bmatrix}.$$

When case II (see Sec. 5.1.3) arises, the point $\mathbf{x} = \mathbf{x}(\lambda)$ is feasible for any $\lambda > 0$, and $\langle \mathbf{c}, \mathbf{x}(\lambda) \rangle \to +\infty$ for $\lambda \to +\infty$.

Since $\mathbf{x}$ is a feasible point, $A\mathbf{x} = A\bar{\mathbf{x}} - \lambda A\mathbf{z} = \mathbf{b}$, and, since $A\bar{\mathbf{x}} = \mathbf{b}$, we have $A\mathbf{z} = \mathbf{0}$ and $\mathbf{z} \geqslant \mathbf{0}$.

Passing to the notation of the problem (5.42) we obtain the system

$$B_j \mathbf{z} = \mathbf{0}, \quad \mathbf{z} \geqslant \mathbf{0}, \tag{5.44}$$

$$\langle A_j^T \mathbf{y} - \mathbf{c}_j, \mathbf{z} \rangle > 0. \tag{5.45}$$

In this case the procedure in an iterative step in the extremal problem consists in the following: the vector $\begin{bmatrix} A_j \mathbf{z} \\ 0 \\ \cdots \\ 0 \end{bmatrix}$ is introduced in the basis of the problem (5.38)-(5.41). We stress that it is the vector $\begin{bmatrix} A_j \mathbf{z} \\ 0 \\ \cdots \\ 0 \end{bmatrix}$ that is introduced and not the vector $\begin{bmatrix} A_j \mathbf{z} \\ \mathbf{c}_j \end{bmatrix}$ because for the latter the condition (5.40) is not fulfilled since the set $\{\mathbf{x}: B_j \mathbf{x} = \mathbf{0}, \mathbf{x} \geqslant \mathbf{0}\}$ is a cone whose any element is representable as a positive linear combination of its edges (see Sec. 2.3.6).

Chapter 6

PENALTY FUNCTION METHOD

6.1. Description of the Method

6.1.1. As a rule, numerical methods of solving unconditional minimization and mathematical programming problems differ from one another qualitatively from the standpoint of the amount of calculations they require. For instance, in addition to difficulties appearing in the gradient descent method (Chap. 9, Sec. 9.5.1) for solving problems of finding unconditional minimum, the conditional gradient method (Chap. 10, Sec. 10.3.1) of solving mathematical programming problems involves the solution of an additional linear programming problem. In this connection much progress can be achieved when attempting to reduce a mathematical programming problem to a sequence of unconditional minimization problems.

6.1.2. The idea of the *penalty function method* consists in the following. The original mathematical programming problem

$$\min_{x \in X} \varphi(\mathbf{x}) = m \tag{6.1}$$

is replaced by the problem of unconditional minimization of a one-parameter function

$$M(\mathbf{x}, \beta) = \varphi(\mathbf{x}) + \frac{1}{\beta} \psi(\mathbf{x})$$

or by the problem of minimizing the function $M(\mathbf{x}, \beta)$ on a set Γ having a simpler structure from the viewpoint of numerical minimization methods than the original set X. The function $\psi(\mathbf{x})$ is chosen so that the solution to the problem

$$\min_{x \in \Gamma} M(\mathbf{x}, \beta) = m(\beta) \tag{6.2}$$

converges for $\beta \to 0$ to the solution of the original problem or, when such an attempt fails, so that at least $m(\beta) \to m$ for $\beta \to 0$.

Thus, the penalty function method consists in constructing the function $M(\mathbf{x}, \beta)$, selecting a sequence $\{\beta_k\}$, and solving the problems of the type of (6.2) corresponding to the sequence.

6.1.3. We consider the representation

$$X = \{\mathbf{x} \in \Gamma: \mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}\},$$

of the set X, where

$$\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})),$$

and the set Γ possesses, as was said above, a simpler structure. As a rule, $\Gamma = E_n$, and then either (6.2) is an unconditional minimization problem or $\Gamma = \{\mathbf{x}: \mathbf{x} \geqslant \mathbf{0}\}$ or $\Gamma = \{\mathbf{x}: \mathbf{a} \leqslant \mathbf{x} \leqslant \mathbf{b}\}$. In the last two cases the complexity of the numerical methods for solving the problem (6.2) is approximately the same as in the case when $\Gamma = E_n$.

6.1.4. Definition. A continuous function

$$\frac{1}{\beta}\psi(\mathbf{x})$$

defined for all $\mathbf{x} \in \Gamma$ and $\beta > 0$ will be called a *penalty* if $\psi(\mathbf{x})$ is such that

$$\begin{aligned} &\psi(\mathbf{x}) = 0 \quad \text{for all} \quad \mathbf{x} \in X, \\ &\psi(\mathbf{x}) > 0 \quad \text{for all} \quad \mathbf{x} \in \Gamma \setminus X. \end{aligned} \tag{6.3}$$

The following form of the function $\psi(\mathbf{x})$ is rather frequently used

$$\psi(\mathbf{x}) = \sum_{i=1}^{m} |\min\{f_i(\mathbf{x}), 0\}|^p, \quad p > 0. \tag{6.4}$$

For this choice of the function $\psi(\mathbf{x})$ the conditions (6.3) are obviously fulfilled.

The penalty function method can be demonstrated visually in the following way. Let

$$X = \{x \in E_1: a \leqslant x \leqslant b\}.$$

Then

$$\psi(x) = [\min\{b - x, 0\}]^p + [\min\{x - a, 0\}]^p, \quad p = 2.$$

Figure 6.1 represents the function $\frac{1}{\beta}\psi(x)$ for two different values of β: $\beta_1 > \beta_2 > 0$.

Let $\varphi(x)$ be a linear function. Figure 6.2 demonstrates the functions $M(x, \beta_1)$ and $M(x, \beta_2)$. The points y_{β_1}

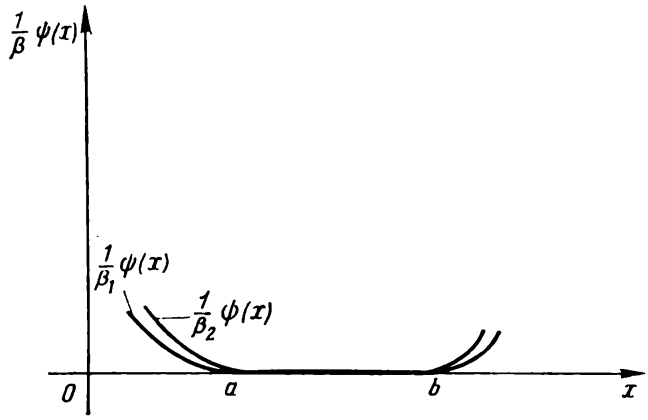


Fig. 6.1

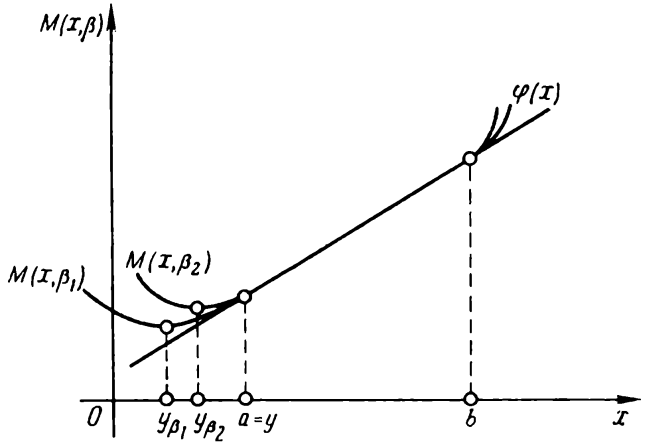


Fig. 6.2

and y_{β_2} are solutions to the corresponding problems of the form (6.2), i.e. the points of absolute minima of the functions $M(x, \beta_1)$ and $M(x, \beta_2)$. By y is denoted the solution to the original problem (6.1), i.e. the point of minimum of the linear function $\varphi(x)$ on the closed interval $[a, b]$. Figure 6.2 shows that the corresponding points y_{β_k} tend to the point y for $\beta_k \to 0$.

Another example is the choice of the function $\psi(\mathbf{x})$ in the form

$$\psi(\mathbf{x}) = \exp\left\{\sum_{i=1}^{m} |\min\{f_i(\mathbf{x}), 0\}|^p\right\} - 1, \quad p > 0.$$

6.1.5. When (6.1) is a convex programming problem, i.e. the function $\varphi(\mathbf{x})$ is convex, the functions $f_i(\mathbf{x})$ are concave, and the set Γ is convex, then the function $\psi(\mathbf{x})$ determined by the formula (6.4) is convex on the set Γ for $p \geqslant 1$, and (6.2) becomes a problem of finding minimum of a convex function $M(\mathbf{x}, \beta)$ on the convex set Γ.

6.1.6. The problem

$$\min_{x \in X} \varphi(\mathbf{x})$$

where

$$X = \{\mathbf{x} \in E_n : \mathbf{f}(\mathbf{x}) \geqslant \mathbf{0}\}$$

is sometimes replaced by the equivalent problem of minimizing $\varphi(\mathbf{x})$ under the conditions $\mathbf{f}(\mathbf{x}) - \mathbf{u} = \mathbf{0}$ and $\mathbf{u} \geqslant \mathbf{0}$, and the quadratic function

$$\frac{1}{\beta}\psi(\mathbf{x}, \mathbf{u}) = \frac{1}{\beta}\|\mathbf{f}(\mathbf{x}) - \mathbf{u}\|^2$$

is taken as a penalty. In this case (6.2) is the problem of minimizing the function

$$M(\mathbf{x}, \mathbf{u}, \beta) = \varphi(\mathbf{x}) + \frac{1}{\beta}\psi(\mathbf{x}, \mathbf{u})$$

on the set $\mathbf{x} \in E_n$, $\mathbf{u} \geqslant \mathbf{0}$.

6.1.7. The choice of the penalty must be compatible with the method of minimizing the function $M(\mathbf{x}, \beta)$, and therefore, as will be seen in the subsequent chapters, one should take into consideration the smoothness of the penalty, the simplicity of the calculation of the values of the function $M(\mathbf{x}, \beta)$ and its derivatives, the properties of convex functions, etc.

We note that for $p = 1$ there can be a loss of smoothness of the function determined by (6.4) in comparison with the case $p = 2$. For instance, for the example in Sec. 6.1.4 it can readily be verified that the function

$$\psi(x) = |\min\{b - x, 0\}| + |\min\{x - a, 0\}|$$

is non-differentiable at the points a and b.

6.1.8. Finally, we elucidate the name of the method. The term "penalty function method" appears quite natural since it explains the essence of the method: when the constraints are violated, i.e. when $\mathbf{x}_1 \in \Gamma \setminus X$, the function $\varphi(\mathbf{x})$ "suffers a penalty" of $\frac{1}{\beta}\psi(\mathbf{x})$.

6.2. Convergence Theorems

6.2.1. A function

$$\Psi(\mathbf{x}, \beta) = \beta\varphi(\mathbf{x}) + \varphi(\mathbf{x})$$

defined on the set Γ for all $\beta > 0$ will be called a *penalty function* of the problem (6.1) if $\psi(\mathbf{x})$ satisfies the conditions (6.3). It is obvious that

$$\Psi(\mathbf{x}, \beta) = \beta M(\mathbf{x}, \beta).$$

We note that this form of a penalty function somewhat disguises its meaning as a penalty for violation of constraints, however it will be seen from what follows that in the numerical realization of minimization methods this form of the function $\Psi(\mathbf{x}, \beta)$ sometimes possesses an advantage as compared to $M(\mathbf{x}, \beta)$.

Let

$$\mathbf{y}_\beta = \arg\min\{\Psi(\mathbf{x}, \beta) \mid \mathbf{x} \in \Gamma\}, \quad Y_\beta = \{\mathbf{y}_\beta\},$$
$$\mathbf{y} = \arg\min\{\varphi(\mathbf{x}) \mid \mathbf{x} \in X\}, \quad Y = \{\mathbf{y}\}.$$

Proposition. If $\mathbf{y}_\beta \in X$, then $\mathbf{y}_\beta \in Y$. Indeed, since $\mathbf{y}_\beta \in X$, we have

$$\min_{\mathbf{x}\in\Gamma} M(\mathbf{x}, \beta) = M(\mathbf{y}_\beta, \beta) = \varphi(\mathbf{y}_\beta)$$
$$\leqslant \varphi(\mathbf{y}) + \frac{1}{\beta}\psi(\mathbf{y}) = \varphi(\mathbf{y}) \leqslant \varphi(\mathbf{y}_\beta),$$

and, consequently, $\varphi(\mathbf{y}) = \varphi(\mathbf{y}_\beta)$, whence $\mathbf{y}_\beta \in Y$. $\triangle$

6.2.2. Theorem. *If there exist* $\mathbf{y}$ *and* $\mathbf{y}_\beta$, *then*

$$\lim_{\beta\to 0} \Psi(\mathbf{y}_\beta, \beta) = 0, \tag{6.5}$$

$$\lim_{\beta\to 0} \psi(\mathbf{y}_\beta) = 0. \tag{6.6}$$

Proof. 1. We shall show that for $\beta_1 \geqslant \beta_2 > 0$ the inequality

$$\frac{1}{\beta_1} \Psi(y_{\beta_1}, \beta_1) \leqslant \frac{1}{\beta_2} \Psi(y_{\beta_2}, \beta_2) \tag{6.7}$$

holds. The fact that

$$\Psi(y_{\beta_1}, \beta_1) = \min_{x \in \Gamma} \Psi(x, \beta_1) \leqslant \Psi(y_{\beta_2}, \beta_1)$$

and the inequality

$$\frac{1}{\beta_1} \psi(y_{\beta_2}) \leqslant \frac{1}{\beta_2} \psi(y_{\beta_2})$$

imply

$$\frac{1}{\beta_1} \Psi(y_{\beta_1}, \beta_1) \leqslant \varphi(y_{\beta_2}) + \frac{1}{\beta_1} \psi(y_{\beta_2})$$

$$\leqslant \varphi(y_{\beta_2}) + \frac{1}{\beta_2} \psi(y_{\beta_2}) = \frac{1}{\beta_2} \Psi(y_{\beta_2}, \beta_2).$$

Further, since $\psi(y) = 0$, we have

$$\Psi(y_\beta, \beta) \leqslant \Psi(y, \beta) = \beta\varphi(y) + \psi(y) = \beta\varphi(y),$$

i.e.

$$\frac{1}{\beta} \Psi(y_\beta, \beta) \leqslant \varphi(y).$$

This and (6.7) (i.e. the monotonicity of $\frac{1}{\beta}\Psi(y_\beta, \beta)$ as a function of β) imply the existence of a finite limit

$$\Psi^* = \lim_{\beta \to 0} \frac{1}{\beta} \Psi(y_\beta, \beta),$$

and hence (6.5) is valid.

2. We first show that

$$\varphi(y_{\beta_1}) \leqslant \varphi(y_{\beta_2}) \tag{6.8}$$

for $\beta_1 > \beta_2 > 0$. The inequalities

$$\Psi(y_{\beta_1}, \beta_1) \leqslant \Psi(y_{\beta_2}, \beta_1),$$
$$\Psi(y_{\beta_2}, \beta_2) \leqslant \Psi(y_{\beta_1}, \beta_2)$$

imply

$$\beta_1\varphi(y_{\beta_1}) + \psi(y_{\beta_1}) \leqslant \beta_1\varphi(y_{\beta_2}) + \psi(y_{\beta_2}),$$
$$\beta_2\varphi(y_{\beta_2}) + \psi(y_{\beta_2}) \leqslant \beta_2\varphi(y_{\beta_1}) + \psi(y_{\beta_1}).$$

Adding these inequalities together we find

$$(\beta_1 - \beta_2)\,[\varphi(\mathbf{y}_{\beta_1}) - \varphi(\mathbf{y}_{\beta_2})] \leqslant 0,$$

and, since $\beta_1 - \beta_2 > 0$, we have

$$\varphi(\mathbf{y}_{\beta_1}) - \varphi(\mathbf{y}_{\beta_2}) \leqslant 0,$$

i.e. (6.8) takes place.

Further, since $\psi(\mathbf{x}) \geqslant 0$,

$$\varphi(\mathbf{y}_\beta) \leqslant \varphi(\mathbf{y}_\beta) + \frac{1}{\beta}\psi(\mathbf{y}_\beta) = \frac{1}{\beta}\Psi(\mathbf{y}_\beta, \beta) \leqslant \Psi^*.$$

This and (6.8) (i.e. the monotonicity of $\varphi(\mathbf{y}_\beta)$) imply the existence of a finite limit

$$\varphi^* = \lim_{\beta \to 0} \varphi(\mathbf{y}_\beta).$$

Therefore (see (6.5)),

$$\lim_{\beta \to 0} \psi(\mathbf{y}_\beta) = \lim_{\beta \to 0} [\Psi(\mathbf{y}_\beta, \beta) - \beta\varphi(\mathbf{y}_\beta)] = 0. \quad \triangle$$

It now becomes clear that to calculate the values of the penalty function it is convenient to use the form

$$\Psi(\mathbf{x}, \beta) = \beta\varphi(\mathbf{x}) + \psi(\mathbf{x})$$

since $\beta\varphi(\mathbf{y}_\beta) \to 0$ and $\psi(\mathbf{y}_\beta) \to 0$, and, therefore, when calculating the values of $\Psi(\mathbf{y}_\beta, \beta)$, magnitudes of approximately the same order are added together.

In the numerical realization of the penalty function method there arise problems of selecting the initial value β_0 of the parameter β and the way β_k is varied. The difficulty consists in the following. The choice of a sufficiently small β_0 allows one to hope that $\mathbf{y}_{\beta_0}$ will be close to $\mathbf{y}$. However, as a rule, the rate of convergence of gradient methods for calculating the points of minimum of the function $\Psi(\mathbf{x}, \beta)$ decreases together with the parameter β whereas the choice of a large β_0 leads to an opposite situation.

6.2.3. In what follows some possibilities of approximate calculation of the point $\mathbf{y}_\beta$ will be provided. We denote $\Psi(\mathbf{y}_\beta, \beta) = \Psi^*(\beta)$ and assume that for each value of the parameter $\beta > 0$ there exists a point $\tilde{\mathbf{y}}_\beta \in \Gamma$ such that

$$\Psi(\tilde{\mathbf{y}}_\beta, \beta) \leqslant \Psi^*(\beta) + \xi(\beta), \quad \xi(\beta) > 0.$$

6.2.4. Theorem. *If there are* $\mathbf{y}$ *and* $\mathbf{y}_\beta$ *and if* $\zeta(\beta) \to 0$ *for* $\beta \to 0$, *then*

$$\lim_{\beta \to 0} \Psi(\tilde{\mathbf{y}}_\beta, \beta) = 0. \tag{6.9}$$

Moreover, if the function $\varphi(\mathbf{x})$ *is bounded from below on the set* $\{\tilde{\mathbf{y}}_\beta\}$ *for all sufficiently small* $\beta > 0$, *then*

$$\lim_{\beta \to 0} \psi(\tilde{\mathbf{y}}_\beta) = 0. \tag{6.10}$$

Proof. The first assertion of the theorem obviously follows from the inequality

$$\Psi(\mathbf{y}_\beta, \beta) \leqslant \Psi(\tilde{\mathbf{y}}_\beta, \beta) \leqslant \Psi(\mathbf{y}_\beta, \beta) + \zeta(\beta)$$

and (6.5).

Since

$$0 \leqslant \psi(\tilde{\mathbf{y}}_\beta) = \Psi(\tilde{\mathbf{y}}_\beta, \beta) - \beta\varphi(\tilde{\mathbf{y}}_\beta),$$

(6.9) and the fact that the function $\varphi(\tilde{\mathbf{y}}_\beta)$ is bounded from below imply (6.10). $\triangle$

6.2.5. Theorem. *If*

(a) $\varphi(\mathbf{x})$, $f_i(\mathbf{x})$ $(i = \overline{1, m})$ *and* $\psi(\mathbf{x})$ *are continuous on* Γ,

(b) *for all* $\beta > 0$ *there exist* $\mathbf{y}$ *and* $\mathbf{y}_\beta$,

(c) *there is a closed and bounded set* $G \subseteq \Gamma$ *such that* $\tilde{\mathbf{y}}_\beta \in G$ *for all sufficiently small* $\beta > 0$,

(d) $\lim\limits_{\beta \to 0} \dfrac{\zeta(\beta)}{\beta} = 0$,

then

$$\lim_{\beta \to 0} \varphi(\tilde{\mathbf{y}}_\beta) = \varphi(\mathbf{y}), \qquad \lim_{\beta \to 0} \rho(\tilde{\mathbf{y}}_\beta, Y) = 0.$$

As before, here $Y = \operatorname{Arg\,min}\{\varphi(\mathbf{x}): \mathbf{x} \in X\}$.

Proof. Assume that the theorem is not true, i.e. there is a number $\delta > 0$ and a sequence $\{\beta_k\} \to 0$, $k \to \infty$, such that the inequality

$$|\varphi(\tilde{\mathbf{y}}_{\beta_k}) - \varphi(\mathbf{y})| \geqslant \delta$$

holds for all $k = 0, 1, \ldots$. Since $\psi(\mathbf{x}) \geqslant 0$ for all $\mathbf{x}$ and $\psi(\mathbf{y}) = 0$, we have

$$\beta_k \varphi(\tilde{\mathbf{y}}_{\beta_k}) \leqslant \beta_k \varphi(\tilde{\mathbf{y}}_{\beta_k}) + \psi(\tilde{\mathbf{y}}_{\beta_k}) = \Psi(\tilde{\mathbf{y}}_{\beta_k}, \beta_k)$$
$$\leqslant \Psi(\mathbf{y}_{\beta_k}, \beta_k) + \zeta(\beta_k) \leqslant \Psi(\mathbf{y}, \beta_k) + \zeta(\beta_k) = \beta_k \varphi(\mathbf{y}) + \zeta(\beta_k). \tag{6.11}$$

Since $\tilde{\mathbf{y}}_{\beta_k} \in G$, we shall assume, without changing the indexing, that

$$\lim_{k \to \infty} \tilde{\mathbf{y}}_{\beta_k} = \tilde{\mathbf{y}}.$$

Then from (6.11) and the condition (d) it follows that

$$\varphi(\tilde{\mathbf{y}}) \leqslant \varphi(\mathbf{y}).$$

However, the continuous function $\varphi(\mathbf{x})$ is bounded from below on the set G (see the condition (c)) for sufficiently small $\beta_k > 0$, and therefore, in view of 6.2.4,

$$\psi(\tilde{\mathbf{y}}) = \lim_{k \to \infty} \psi(\tilde{\mathbf{y}}_{\beta_k}) = 0.$$

Consequently, $\tilde{\mathbf{y}} \in X$, whence

$$\varphi(\tilde{\mathbf{y}}) = \varphi(\mathbf{y}),$$

which contradicts the assumption.

From what was said it is clearly seen that any limiting point of the family $\{\tilde{\mathbf{y}}_\beta\}$ belongs to the set Y.

Assume that the assertion that $\rho(\tilde{\mathbf{y}}_\beta, Y) \to 0$ for $\beta \to 0$ is false. Then there are a number $\varepsilon > 0$ and a subsequence $\{\tilde{\mathbf{y}}_{\beta_k}\} \subset \{\tilde{y}_\beta\}$ such that $\rho(\tilde{\mathbf{y}}_{\beta_k}, Y) > \varepsilon$ for all $k = 0, 1, \ldots$. With regard to the condition (c), we can assume, without loss of generality, that the sequence $\{\tilde{\mathbf{y}}_{\beta_k}\}$ converges to a point $\tilde{\mathbf{y}}$. However, as was seen, $\tilde{\mathbf{y}} \in Y$, and hence the relation $\lim_{k \to \infty} \rho(\tilde{\mathbf{y}}_{\beta_k}, Y) = \lim_{k \to \infty} \rho(\tilde{\mathbf{y}}_{\beta_k}, \tilde{\mathbf{y}}) = 0$ holds, which contradicts the assumption. $\triangle$

Corollary 1. *If* $\{\tilde{y}_\beta\} \in \Gamma \setminus X$, *then any limiting point* $\tilde{y}$ *belongs to the intersection of the boundary* $G(Y)$ *of the set* Y *and the boundary* $G(X)$ *of the set* X.

Indeed, since $\rho(\tilde{y}_\beta, Y) \to 0$ for $\beta \to 0$, any limiting point $\tilde{y}$ of the family $\{\tilde{y}_\beta\}$ belongs to the set Y. However, $\{\tilde{y}_\beta\} \notin Y$, and, consequently, $\tilde{y}$ does not belong to the interior of the set Y, and therefore $\tilde{y} \in G(Y)$. It is similarly proved that $\tilde{y} \in G(X)$. $\triangle$

Corollary 2. *If* $Y \subset \operatorname{int} X$, *then there is an index* l *such that* $y_{\beta_l} \in Y$.

Indeed, if otherwise, then (see Sec. 6.2.1) $\{y_{\beta_k}\} \subset \Gamma \setminus X$, and any limiting point $\bar{y}$ belongs to $G(Y) \cap G(X)$; consequently, $G(Y) \cap G(X) \neq \varnothing$, which is a contradiction. $\triangle$

Remark. If $\varphi(x)$ and $\psi(x)$ are convex and $Y \subset \operatorname{int} X$, then $y_\beta \in Y$ for any $\beta > 0$ since the convexity of $M(x, \beta)$ and the condition $Y \subset \operatorname{int} X$ imply that $Y = \operatorname{Arg\,min}\{M(x, \beta): x \in \Gamma\}$.

6.2.6. On the rate of convergence of the penalty function method. Denote $f(x) = \min\limits_{i=\overline{1,m}} f_i(x)$. We assume that

(a) $\varphi(x)$ and $f_i(x)$ $(i = \overline{1, m})$ are continuous on a convex and closed set Γ,

(b) for all $\beta > 0$ there exist y and y_β,

(c) there is a compactum $G \subset \Gamma$ such that $y_\beta \in G$ for sufficiently small $\beta > 0$,

(d) $\psi(x) = \sum\limits_{i=1}^{m} |\min\{f_i(x), 0\}|^p, \quad p > 0,$

(e) $|\varphi(x_1) - \varphi(x_2)| \leqslant L\rho(x_1, x_2)$ for any $x_1, x_2 \in \Gamma$,

(f) there are $\delta > 0$ and $\alpha > 0$ such that $|f(x)| \geqslant \alpha\rho(x, X)$ for all $x \in U_\delta(X) \setminus X$, where $U_\delta(X) = \{x \in \Gamma: \rho(x, X) \leqslant \delta\}$.

Denote $\Delta(\beta) = \varphi(y) - \frac{1}{\beta}\Psi(y_\beta, \beta)$.

6.2.7. Theorem. *If the conditions* (a) *to* (f) *are fulfilled, then for* $p \leqslant 1$ *there is a sufficiently small* $\beta = \bar{\beta}$ *such that*

$\Delta(\beta) = 0$ *for all* $\beta \in (0, \bar{\beta}]$ *and for* $p > 1$ *the inequality*

$$0 \leqslant \Delta(\beta) \leqslant C\beta^{1/(p-1)}, \quad C = L\left(\frac{L}{p\alpha^p}\right)^{1/(p-1)}\left(1 - \frac{1}{p}\right)$$

holds for all sufficiently small $\beta > 0$.

Proof. If $\mathbf{y}_\beta \in X$, then (see Sec. 6.2.1) $\mathbf{y}_\beta \in Y$, whence $\Delta(\beta) = 0$.

Let $\mathbf{y}_\beta \in \Gamma \setminus X$. Since $\mathbf{y}_\beta \in G$ and G is a compactum, Theorem 6.2.5 implies that $\rho(\mathbf{y}_\beta, X) \to 0$ for $\beta \to 0$ (see Sec. 6.2.5, Corollary (1)). Therefore, $\mathbf{y}_\beta \in U_\delta(X) \setminus X$ for sufficiently small values of β. By the hypothesis, the condition (f) holds for these values of $\mathbf{y}_\beta$.

Denote by $\mathbf{v}_\beta$ the projection of the point $\mathbf{y}_\beta$ on the set X: $\rho(\mathbf{y}_\beta, X) = \rho(\mathbf{y}_\beta, \mathbf{v}_\beta)$. We note that

$$\sum_{i=1}^{m} |\min\{f_i(\mathbf{y}_\beta), 0\}|^p \geqslant |\min_{i=\overline{1,m}} f_i(\mathbf{y}_\beta)|^p = |f(\mathbf{y}_\beta)|^p.$$

Further,

$$\frac{1}{\beta}\Psi(\mathbf{y}_\beta, \beta) = \varphi(\mathbf{y}_\beta) + \frac{1}{\beta}\sum_{i=1}^{m} |\min\{f_i(\mathbf{y}_\beta), 0\}|^p$$

$$\geqslant \varphi(\mathbf{v}_\beta) - |\varphi(\mathbf{v}_\beta) - \varphi(\mathbf{y}_\beta)| + \frac{1}{\beta}|f(\mathbf{y}_\beta)|^p \geqslant \varphi(\mathbf{y})$$

$$- L\rho(\mathbf{y}_\beta, \mathbf{v}_\beta) + \frac{\alpha^p}{\beta}[\rho(\mathbf{y}_\beta, X)]^p = \varphi(\mathbf{y}) - L\rho + \frac{\alpha^p}{\beta}\rho^p.$$

Here, $\rho \triangleq \rho(\mathbf{y}_\beta, X)$. Thus, $\Delta(\beta) \leqslant \rho\left(L - \frac{\alpha^p}{\beta}\rho^{p-1}\right)$.

Since $\rho \to 0$ for $\beta \to 0$ (see 6.2.5), we have $\rho^{p-1} \to +\infty$ for $0 < p < 1$, and, consequently, there exists a sufficiently small $\bar{\beta}$, such that $\Delta(\beta) \leqslant 0$ for any $\beta \in (0, \bar{\beta}]$, and hence $\Delta(\beta) = 0$.

If $p = 1$, then $\Delta(\beta) \leqslant \rho\left(L - \frac{\alpha}{\beta}\right)$, and therefore $\Delta(\beta) = 0$ for $\beta \leqslant \alpha/L$. Finally, for $p > 1$ the inequality

$$\Delta(\beta) \leqslant \max_{\rho \geqslant 0}\left(L\rho - \frac{\alpha^p}{\beta}\rho^p\right) = C\beta^{1/(p-1)}$$

is fulfilled since the maximum is attained at $\rho = \left(\frac{L}{p\alpha^p}\beta\right)^{1/(p-1)}$. $\triangle$

Remark. If the exponential function

$$\psi(x) = \exp\left\{\sum_{i=1}^{m} |\min\{f_i(x), 0\}|^p\right\} - 1$$

is taken as $\psi(x)$, then the inequality $e^y \geqslant 1 + y$ implies the validity of Theorem 6.2.7 for this case as well.

6.2.8. Problem solving practice indicates that when the widely applied penalties of the type (6.4) are used, the convergence is substantially decelerated with decreasing β, and, besides, the effect of rounding errors strengthens at each step of the unconditional minimization.

Such a situation can be caused by a number of factors, e.g. it can appear since a penalty of the form (6.4) does not take into account specific features of the structure of the feasible region owing to the ravine shape of the function $M(x, \beta)$, etc.

As a rule, the techniques improving the situation to a certain extent, i.e. making it possible to accelerate the process of determining the point y to a given degree of accuracy, are of a heuristic character and are based on specialist's experience and intuition. Attempts are also made to elaborate standard techniques meant for improving certain characteristics of the method.

The following form of a penalty function guarantees the finiteness of the method for any $p > 0$:

$$F(x, \beta) = \varphi(x) + \frac{1}{\beta}\sum_{i=1}^{m} |\min\{f_i(x), 0\}|^p + \frac{1}{\beta}\sum_{i=1}^{m} |\min\{f_i(x), 0\}|^{1/p}. \quad (6.12)$$

It follows from the proof of Theorem 6.2.7 that

$$0 \leqslant \Delta(\beta) = \varphi(y) - F(y_\beta, \beta) \leqslant \rho\left[L - \frac{1}{\beta}(\alpha^p \rho^{p-1} + \alpha^{1/p}\rho^{1/p-1})\right].$$

For $p = 1$ we have $\Delta(\beta) \leqslant \rho\left(L - \frac{2\alpha}{\beta}\right)$, and, consequently, $\Delta(\beta) = 0$ for sufficiently small β.

For $p \neq 1$ we obviously have

$$\frac{1}{\beta}(\alpha^p \rho^{p-1} + \alpha^{1/p}\rho^{1/p-1}) \to +\infty \quad \text{for} \quad \beta \to 0.$$

Since $\rho \to 0$ for $\beta \to 0$, for small β we again arrive at a situation when $\Delta(\beta) = 0$.

Thus, for any $p > 0$ there is $\bar{\beta}$ such that $\Delta(\beta) = 0$ for all $\beta \in (0, \bar{\beta}]$. Here it can be expected that the farther p lies from 1 (either in the direction towards zero or towards $+\infty$), the earlier the situation where $\Delta(\beta) = 0$ arises.

However, to investigate this question more thoroughly, it is necessary to have more exact estimates that require taking into consideration the properties of the functions $f_i(\mathbf{x})$ and the fact that the point $\mathbf{v}_\beta$ satisfies the corresponding constraints $f_i(\mathbf{x}) \geqslant 0$.

6.2.9. γ-majorizability. As was said in Sec. 6.2.7, the convergence of the penalty function method is strongly affected by the properties of the functions $f_i(\mathbf{x})$. When deriving the estimate for $\Delta(\beta)$ in Sec. 6.2.7 we used the property (f). However, it remained unclear how the fulfilment of the condition (f) could be verified for a specific set X. We shall show that in some cases this condition is sure to hold.

Theorem. *If the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are concave and continuous on a convex, closed, and bounded set Γ and the set $X = \{\mathbf{x} \in \Gamma: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ is regular in Slater's sense, then there are $\delta, \alpha > 0$, such that*

$$|f(\mathbf{x})| = \left| \min_{i=\overline{1, m}} f_i(\mathbf{x}) \right| \geqslant \alpha \rho(\mathbf{x}, X)$$

for all $\mathbf{x} \in U_\delta(X) \setminus X$.

Proof. Since the set X is regular in Slater's sense, (see Sec. 3.4.3.), there exists a point $\tilde{x} \in X$ and a number $\lambda > 0$ such that $f_i(\tilde{\mathbf{x}}) \geqslant \lambda$ $(i = \overline{1, m})$. Let $\mathbf{x} \in \Gamma \setminus X$, i.e. $f(x) = \min\limits_{i=\overline{1, m}} f_i(\mathbf{x}) < 0$. On the closed interval $[\tilde{\mathbf{x}}, \mathbf{x}]$ there is a point $\mathbf{z}$ such that $f(\mathbf{z}) = 0$. The concave function $f(\mathbf{x})$ satisfies the inequality (see Sec. 2.4.14)

$$\frac{f(\mathbf{z}) - f(\mathbf{x})}{\|\mathbf{z} - \mathbf{x}\|} \geqslant \frac{f(\tilde{\mathbf{x}}) - f(\mathbf{z})}{\|\tilde{\mathbf{x}} - \mathbf{z}\|},$$

and, since $\|\mathbf{z} - \tilde{\mathbf{x}}\| \leqslant r = \operatorname{diam} \Gamma < +\infty$, we have

$$f(\mathbf{x}) \leqslant -\frac{\|\mathbf{z} - \mathbf{x}\|}{\|\tilde{\mathbf{x}} - \mathbf{z}\|} f(\tilde{\mathbf{x}}) \leqslant -\frac{\lambda}{r} \|\mathbf{z} - \mathbf{x}\| \leqslant -\frac{\lambda}{r} \rho(\mathbf{x}, X).$$

Hence, the condition (f) is fulfilled for $\alpha = \lambda/r$.

It can also be easily proved that the condition (f) holds when the functions $f_i(\mathbf{x})$ are linear. In this case even the regularity of the set X is not needed.

It often happens that the inequality $|f(\mathbf{x})| \geqslant \alpha\rho(\mathbf{x}, X)$ is not fulfilled while for some $\gamma > 0$ the condition $|f(\mathbf{x})| \geqslant \alpha[\rho(\mathbf{x}, X)]^{\gamma}$ holds.

Definition. We shall say that the set $X = \{\mathbf{x} \in \Gamma: f_i(\mathbf{x}) \geqslant 0, i = \overline{1, m}\}$ is γ-*majorizable* if

$$|f(\mathbf{x})| \geqslant \alpha[\rho(\mathbf{x}, X)]^{\gamma}, \quad \alpha > 0, \quad \gamma > 0$$

for all $\mathbf{x} \in U_{\delta}(X) \setminus X$.

For example, if $n = 1$, $m = 2$, $f_1(x) = -x^3$, and $f_2(x) = (x+1)^3$, then the feasible region $X = [-1, 0]$ is γ-majorizable for $\gamma = 3$.

If the exponent γ is known for a specific set X, then, when choosing a penalty in the form (6.4), the estimate for the error $\Delta(\beta)$ can obviously be written in the following way:

$$0 \leqslant \Delta(\beta) \leqslant \begin{cases} \delta_1(\beta) = C\beta^{1/(\gamma p - 1)} & \text{for } \gamma p > 1, \\ 0 & \text{for } \gamma p \leqslant 1, \end{cases} \tag{6.13}$$

and in the case when $\gamma p \leqslant 1$ the condition that the set Γ is compact is important. It is clear that, for instance, for $\gamma = 2$ and $p = 1$ the application of a penalty in the form (6.4) does not guarantee the finiteness of the procedure. Only when $0 < p \leqslant 1/2$, i.e. when a penalty with a fractional exponent is used, the number of steps with respect to β is finite.

If a penalty with a fractional exponent $1/p$, $p > 1$, is used, the parameter $\Delta(\beta)$ satisfies the estimate

$$0 \leqslant \Delta(\beta) \leqslant \begin{cases} \delta_2(\beta) = C\beta^{p/(\gamma - p)} & \text{for } \gamma > p > 1, \\ 0 & \text{for } \gamma \leqslant p. \end{cases}$$

We note that δ_1 and δ_2 are related by

$$\delta_2 = \delta_1^{p(\gamma p - 1)/(\gamma - p)},$$

and for $\gamma > p > 1$ the parameter

$$\zeta = \frac{p(\gamma p - 1)}{\gamma - p}$$

exceeds unity. Thus, the relation $\delta_2 = \delta_1^{\zeta}$, $\zeta > 1$, characterizes the rate of decrease of $\Delta(\beta)$ for $\gamma > p > 1$ as favourable for the selection of a penalty with a fractional exponent. At the same time, the application of such a penalty leads to a loss of convexity of the function $\Psi(\mathbf{x}, \beta)$ and may contradict the assumption that the points $\mathbf{y}_\beta$ exist. For instance, for $\gamma = 1$ and $p = 1/2$ the inequality

$$\Delta(\beta) \leqslant \sup_{\rho > 0} \left(L\rho - \frac{1}{\beta} \sqrt{\alpha\rho} \right) = +\infty$$

contains no meaningful information, and, in particular, if the one-dimensional problem $\min(-x)$ is solved under the condition $-x \geqslant 0$, then

$$M(x, \beta) = -x + \frac{1}{\beta} \sqrt{|\min\{-x, 0\}|}$$

$$\text{and} \quad \inf M(x, \beta) = -\infty.$$

We see that, generally, in the case of γ-majorizability the application of a penalty of the form (6.4) is often inappropriate for both $p > 1$ and $0 < p \leqslant 1$. In view of this, it is advisable to use a combination of these penalties having the form (6.12), which, in a sense, is universal; namely, it retains the advantages of the two penalties while some of their shortcomings are eliminated. It can be shown that under some rather general assumptions (the γ-majorizability, the concavity of $f_i(\mathbf{x})$, etc.) the function $F(\mathbf{x}, \beta)$ is unimodal for small β.

It should be noted that, practically, in the general case an attempt to find the exponent γ fails, in view of which even the application of a penalty (6.12) does not guarantee the finiteness of the method with respect to β, which is seen from the inequality

$$0 \leqslant \Delta(\beta) \leqslant \rho \left[L - \frac{1}{\beta} \left(\alpha^p \rho^{\gamma p - 1} + \alpha^{1/p} \rho^{\gamma/p - 1} \right) \right], \quad p > 1.$$

Chapter 7

STABILITY PROBLEMS IN MATHEMATICAL PROGRAMMING

7.1. Correctly and Incorrectly Posed Problems

7.1.1. In a great number of practical problems whose mathematical models are mathematical programming problems the initial information is of approximate character. Therefore, there arises the problem of separating out classes of problems and numerical minimization methods for which small disturbances in the initial information are inessential, i.e. produce no strong effect on the solution.

7.1.2. The exact problem. Consider the problem of finding elements $\mathbf{y} \in X$ such that

$$\varphi(\mathbf{y}) = \min_{\mathbf{x} \in X} \varphi(\mathbf{x}),$$

where

$$X = \{\mathbf{x} \in \Gamma:\ \mathbf{f}(\mathbf{x}) \geqslant 0\},$$

$\mathbf{f}(\mathbf{x})$ is a vector function:

$$\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}),\ f_2(\mathbf{x}),\ \ldots,\ f_m(\mathbf{x})),$$

and $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are given continuous scalar functions.

7.1.3. The initial information. In practice the information about $\varphi(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ is of approximate character, i.e. instead of $\varphi(\mathbf{x})$ and $\mathbf{f}(\mathbf{x})$ we know some functions $\varphi_\varepsilon(\mathbf{x})$ and $\mathbf{f}_\varepsilon(\mathbf{x})$ on the set Γ such that

$$\varphi_\varepsilon(\mathbf{x}),\ \mathbf{f}_\varepsilon(\mathbf{x}) \in P_\varepsilon(\varphi,\ \mathbf{f}) = \{\varphi_\varepsilon(\mathbf{x}),\ \mathbf{f}_\varepsilon(\mathbf{x}):\ |\varphi(\mathbf{x}) - \varphi_\varepsilon(\mathbf{x})| \leqslant \varepsilon,\ \|\mathbf{f}(\mathbf{x}) - \mathbf{f}_\varepsilon(\mathbf{x})\| \leqslant \varepsilon,\quad \mathbf{x} \in \Gamma\}.$$

Here the choice of the functions $\varphi_\varepsilon(\mathbf{x})$ and $\mathbf{f}_\varepsilon(\mathbf{x})$ from the set P_ε does not depend on our desire. Hence, the

solution y to the original problem is judged by the solution y_ε to an approximate problem.

7.1.4. The approximate problem. $\varphi_\varepsilon(y) = \min\limits_{x \in X_\varepsilon} \varphi_\varepsilon(x)$ where

$$X_\varepsilon = \{x \in \Gamma:\ f_\varepsilon(x) \geqslant 0\}.$$

Let

$$Y = \{y \in X:\ \varphi(y) = \min_{x \in X} \varphi(x)\} \neq \varnothing,$$

$$Y_\varepsilon = \{y_\varepsilon \in X_\varepsilon:\ \varphi_\varepsilon(y_\varepsilon) = \min_{x \in X_\varepsilon} \varphi_\varepsilon(x)\} \neq \varnothing.$$

7.1.5. Correctness. We say that the problem 7.1.2 is *weakly correct* if

$$\lim_{\varepsilon \to 0} \sup_{y_\varepsilon \in Y} \inf_{y \in Y} \| y_\varepsilon - y \| = 0. \tag{7.1}$$

The problem 7.1.2 will be called *correct* (*correctly posed*) if

$$\lim_{\varepsilon \to 0} \sup_{z', z'' \in Z} \| z' - z'' \| = 0, \tag{7.2}$$

where $Z = Y_\varepsilon \cup Y$.

We elucidate these definitions. The limiting relation (7.1) implies that for $\varepsilon \to 0$ the family of sets Y_ε contracts to a set belonging to Y. In the "ε, δ" language this means that, given any $\delta > 0$, there is $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ and any $y_\varepsilon \in Y_\varepsilon$ there is $y \in Y$ such that $\| y_\varepsilon - y \| \leqslant \delta$. It follows that for any $\delta > 0$ there is $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ and $\varphi_\varepsilon \in P_\varepsilon$ we have

$$| \varphi_\varepsilon(y_\varepsilon) - \varphi(y) | \leqslant \delta.$$

(We suggest that the reader should verify the validity of this assertion.) Thus, in essence, weak correctness implies correctness relative to the functional

$$\lim_{\varepsilon \to 0} \varphi_\varepsilon(y_\varepsilon) = \varphi(y).$$

The condition (7.2) implies that for $\varepsilon \to 0$ the family of sets Y_ε contracts to a set Y consisting of a single element y. Hence, for any $\delta > 0$ there is $\varepsilon_0 > 0$ such that for any $\varepsilon \in (0, \varepsilon_0]$ and any $y_\varepsilon \in Y_\varepsilon$ we have

$$\| y_\varepsilon - y \| \leqslant \delta.$$

It is obvious that a correct problem is always weakly correct but not vice versa.

We shall say that the problem 7.1.2 is *incorrect* (*incorrectly posed*) if the condition (7.2) is violated.

In the case of an incorrectly posed problem 7.1.2, given $\varepsilon > 0$, there always are pairs φ_ε, $\mathbf{f}_\varepsilon$ and $\tilde{\varphi}_\varepsilon$, $\tilde{\mathbf{f}}_\varepsilon$ belonging to the set $P_\varepsilon(\varphi, \mathbf{f})$ for which the solutions $\mathbf{y}_\varepsilon$ and $\tilde{\mathbf{y}}_\varepsilon$ to the corresponding problems of the type 7.1.4 are substantially different and therefore contain no sufficient information about the solution to the original problem.

7.1.6. Examples. *Example 1.* Consider the following pair of problems in E_2:

$$\begin{gathered} \min(-x_1 - x_2), \\ x_1 + x_2 \leqslant 1, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0 \end{gathered} \tag{7.3}$$

and

$$\begin{gathered} \min(-x_1 - x_2), \\ (1 + |\varepsilon| + \varepsilon) x_1 + (1 + |\varepsilon| - \varepsilon) x_2 \leqslant 1, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{gathered}$$

The problem (7.3) is obviously weakly correct and, at the same time, incorrect.

Example 2.

$$\begin{gathered} \min x_2, \\ x_1 + x_2 = 1, \\ (1 + |\varepsilon| + \varepsilon) x_1 + (1 + |\varepsilon| - \varepsilon) x_2 = 1, \\ x_1 \geqslant 0, \quad x_2 \geqslant 0. \end{gathered} \tag{7.4}$$

It is obvious that for $\varepsilon = 0$ the problem (7.4) is incorrect and, moreover, it is not weakly correct.

Basic methods of solving mathematical programming problems were elaborated under a tacit assumption that the problems in question were correct. For instance, the procedure of the simplex method of solving linear programming problems

$$\min_{\mathbf{x} \in R_0} \langle \mathbf{c}, \mathbf{x} \rangle,$$

$$R_0 = \{\mathbf{x}: A\mathbf{x} = b, \quad \mathbf{x} \geqslant \mathbf{0}\}$$

can be interpreted in the following way: at each iteration of the method the elements of the matrix A and the components of the vectors $\mathbf{b}$ and $\mathbf{c}$ are turned into A', $\mathbf{b}'$, and $\mathbf{c}'$, respectively, by identical transformation formulas (Gauss-Jordan elimination formulas), and, if all the transformations are carried out exactly, the iteration results in the appearance of a new problem

$$\min_{\mathbf{x}\in R_0'} \langle \mathbf{c}', \mathbf{x}\rangle,$$

$$R_0' = \{\mathbf{x}: A'\mathbf{x} = \mathbf{b}', \quad \mathbf{x} \geqslant 0\}$$

equivalent to the original one. In practical calculations rounding is always made, and therefore, in reality, an iterative step results not in an equivalent problem but in a new approximate problem. It is clear that in the case of an incorrect problem the solution of the approximate problem may radically differ from that of the original problem. In recent years mathematicians have turned to elaboration of methods meant for solving incorrectly posed problems. One of the main methods of this kind is the so-called *regularization method* which is based on the idea of approximating incorrect problems with correct ones. Therefore it is necessary to separate out classes of correct problems.

7.2. A Class of Correct Problems

7.2.1. We consider the problem 7.1.2 and the problem

$$\min_{\mathbf{x}\in X} \varphi_\varepsilon(\mathbf{x}).$$

We stress that, in contrast to the problem 7.1.4, the feasible region in the approximate problem under study coincides with that in the exact problem, and therefore here we have

$$Y_\varepsilon = \{\mathbf{y}_\varepsilon \in \Gamma: \ \varphi_\varepsilon(\mathbf{y}_\varepsilon) = \min_{\mathbf{x}\in X} \varphi_\varepsilon(\mathbf{x})\}.$$

7.2.2. Naturally, the accuracy in calculating the point $\mathbf{y}_\varepsilon = \arg\min \{\varphi_\varepsilon(\mathbf{x}): \mathbf{x} \in X\}$ should be compatible with the accuracy in setting the objective function $\varphi(\mathbf{x})$, i.e. with the parameter ε. We shall assume that

for each $\varepsilon > 0$ there exists a point $\tilde{y}_\varepsilon \in X$ such that

$$\varphi_\varepsilon(\tilde{y}_\varepsilon) \leqslant \varphi_\varepsilon(y_\varepsilon) + \zeta(\varepsilon), \quad \zeta(\varepsilon) > 0.$$

7.2.3. Theorem. *If*

(a) *the function* $\varphi(x)$ *is strongly convex and continuous on a convex and closed set* X,

(b) $\varphi_\varepsilon(x)$ *is continuous on* X *and*

$$\varphi_\varepsilon(x) \in P_\varepsilon(\varphi) = \{\varphi_\varepsilon(x): \ |\varphi_\varepsilon(x) - \varphi(x)| \leqslant \varepsilon, \quad x \in X\},$$

(c) $\lim\limits_{\varepsilon \to 0} \zeta(\varepsilon) = 0$,

then

$$\lim_{\varepsilon \to 0} \tilde{y}_\varepsilon = y,$$

and, consequently, the problem 7.1.2 *is correct.*

Proof. Since $\varphi_\varepsilon(x) \geqslant \varphi(x) - \varepsilon$ and $\varphi(x) \to +\infty$ for $\|x\| \to \infty$, $x \in X$, we have $\varphi_\varepsilon(x) \to +\infty$; therefore there is $C = \text{const}$ such that the nonempty set $\{x \in X: \ \varphi_\varepsilon(x) \leqslant C\}$ is bounded, which implies that $Y_\varepsilon \neq \varnothing$.

Since $\varphi_\varepsilon(y_\varepsilon) \leqslant \varphi_\varepsilon(y)$,

$$\begin{aligned}\varphi(\tilde{y}_\varepsilon) - \varphi(y) &= [\varphi(\tilde{y}_\varepsilon) - \varphi_\varepsilon(\tilde{y}_\varepsilon)] + [\varphi_\varepsilon(\tilde{y}_\varepsilon) - \varphi(y)] \\ &\leqslant [\varphi(\tilde{y}_\varepsilon) - \varphi_\varepsilon(\tilde{y}_\varepsilon)] + [\varphi_\varepsilon(y_\varepsilon) - \varphi(y)] + \zeta(\varepsilon) \\ &\leqslant [\varphi(\tilde{y}_\varepsilon) - \varphi_\varepsilon(\tilde{y}_\varepsilon)] + [\varphi_\varepsilon(y) - \varphi(y)] + \zeta(\varepsilon) \leqslant 2\varepsilon + \zeta(\varepsilon).\end{aligned}$$

However, in view of the strong convexity of the function $\varphi(x)$, the inequality (see 2.5.5)

$$\|\tilde{y}_\varepsilon - y\|^2 \leqslant \frac{2}{\rho}[\varphi(\tilde{y}_\varepsilon) - \varphi(y)]$$

must hold, and, hence

$$\|\tilde{y}_\varepsilon - y\|^2 \leqslant \frac{4\varepsilon + 2\zeta(\varepsilon)}{\rho}. \quad \triangle$$

7.3. Problems with Exact Constraints. Regularization Method

7.3.1. A.N. Tikhonov suggested the regularization method of solving incorrectly posed problems (including linear programming problems). In this section this method is considered in application to nonlinear programming problems.

The majority of real problems that are formalized as mathematical programming problems contain additional information making it possible to select an element in the solution set Y of the problem 7.1.2, which satisfies some additional conditions; for instance it can be the closest element to a fixed point $\mathbf{x}^0$. In operative planning problems where $\mathbf{y}$ is an optimal plan the element $\mathbf{x}^0$ usually characterises the plan for the foregoing period, a substantial deviation from which may be connected with reorganization and technological reconstruction and thus may lead to additional expenditure.

7.3.2. The normal solution. There naturally arises the following statement of the problem: given $\mathbf{x}^0 \in \Gamma$, it is required to find $\mathbf{y}^0 \in Y$ such that

$$\| \mathbf{y}^0 - \mathbf{x}^0 \|^2 = \min_{\mathbf{y} \in Y} \| \mathbf{y} - \mathbf{x}^0 \|^2. \tag{7.5}$$

The element $\mathbf{y}^0$ is called the *normal solution* to the problem 7.1.2 and the problem (7.5) itself is referred to as the *problem on the normal solution.*

7.3.3. Regularization. The regularization procedure of finding the normal solution to the problem 7.1.2 for a special case when only the objective function is perturbed (while the information about the set X is assumed to be exact) consists in the following.

A regularizing function

$$N_\alpha(\mathbf{x}) = \varphi(\mathbf{x}) + \alpha\Omega(\mathbf{x}), \ \alpha > 0,$$

is introduced, where

$$\Omega(\mathbf{x}) = \| \mathbf{x} - \mathbf{x}^0 \|^2,$$

and the problem of finding an element

$$\mathbf{y}_\alpha = \arg\min \{N_\alpha(\mathbf{x}): \mathbf{x} \in X\}$$

is considered.

7.3.4. In what follows we shall provide a possibility of approximate calculation of the point $\mathbf{y}_\alpha$.

We denote $N_\alpha^* = N_\alpha(\mathbf{y}_\alpha)$ and assume that for each value of the parameter $\alpha > 0$ there exists a point $\tilde{\mathbf{y}}_\alpha \in X$ such that

$$N_\alpha(\tilde{\mathbf{y}}_\alpha) \leqslant N_\alpha^* + \xi(\alpha), \quad \xi(\alpha) > 0.$$

7.4. Convergence

7.4.1. Theorem. *If the function $\varphi(\mathbf{x})$ is continuous on the closed set X and the set Y is nonempty, then, under the condition $\lim\limits_{\alpha \to 0} \frac{\xi(\alpha)}{\alpha} = 0$, there hold*

$$\lim_{\alpha \to 0} \varphi(\tilde{\mathbf{y}}_\alpha) = \varphi(\mathbf{y}) \text{ and } \lim_{\alpha \to 0} \rho(\tilde{\mathbf{y}}_\alpha, Y) = 0.$$

Moreover, if the function $\varphi(\mathbf{x})$ and the set X are convex, then

$$\lim_{\alpha \to 0} \tilde{\mathbf{y}}_\alpha = \mathbf{y}^0.$$

Proof. Since $\varphi(\mathbf{x}) \geqslant \varphi(\mathbf{y})$ for all $\mathbf{x} \in X$, we have $N_\alpha(\mathbf{x}) \to +\infty$ for $\|\mathbf{x}\| \to \infty$ $(\mathbf{x} \in X)$, and, consequently, the nonempty closed set $\{\mathbf{x} \in X: N_\alpha(\mathbf{x}) \leqslant C\}$ is bounded and, by virtue of the continuity of $N_\alpha(\mathbf{x})$, there is $\mathbf{y}_\alpha$. Since

$$\begin{aligned}\varphi(\mathbf{y}^0) \leqslant \varphi(\tilde{\mathbf{y}}_\alpha) \leqslant \varphi(\tilde{\mathbf{y}}_\alpha) + \alpha\Omega(\tilde{\mathbf{y}}_\alpha) \leqslant N_\alpha^* + \xi(\alpha) \\ = \varphi(\mathbf{y}_\alpha) + \alpha\Omega(\mathbf{y}_\alpha) + \xi(\alpha) \leqslant \varphi(\mathbf{y}^0) + \alpha\Omega(\mathbf{y}^0) + \xi(\alpha) \\ \leqslant \varphi(\tilde{\mathbf{y}}_\alpha) + \alpha\Omega(\mathbf{y}^0) + \xi(\alpha),\end{aligned}$$

we have

$$\begin{gathered}\Omega(\tilde{\mathbf{y}}_\alpha) \leqslant \Omega(\mathbf{y}^0) + \frac{\xi(\alpha)}{\alpha}, \\ \varphi(\mathbf{y}^0) \leqslant \varphi(\tilde{\mathbf{y}}_\alpha) \leqslant \varphi(\mathbf{y}^0) + \alpha\Omega(\mathbf{y}^0) + \xi(\alpha).\end{gathered} \tag{7.6}$$

Passing to the limit for $\alpha \to 0$ in the last inequality we obtain

$$\lim_{\alpha \to 0} \varphi(\tilde{\mathbf{y}}_\alpha) = \varphi(\mathbf{y}^0).$$

Since it follows from (7.6) that for all sufficiently small α the set $\{\tilde{y}_\alpha\}$ is bounded, we have

$$\lim_{\alpha \to 0} \rho(\tilde{y}_\alpha, Y) = 0.$$

Finally, let us assume that $\varphi(x)$ and X are convex, then the set Y is also convex. Let $\bar{y}$ be an arbitrary limiting point of the family $\{\tilde{y}_\alpha\}$:

$$\bar{y} = \lim_{k \to \infty} \tilde{y}_{\alpha_k}.$$

As has been proved,

$$\lim_{k \to \infty} \varphi(\tilde{y}_{\alpha_k}) = \varphi(y^0) = \varphi(\bar{y}),$$

and therefore $\bar{y} \in Y$. It follows from (7.6) that

$$\lim_{k \to \infty} \Omega(\tilde{y}_{\alpha_k}) = \Omega(\bar{y}) \leqslant \Omega(y^0),$$

i.e.

$$\min_{y \in Y} \Omega(y) = \Omega(y^0) = \Omega(\bar{y}).$$

However, the strongly convex function $\Omega(x)$ on the convex and closed set Y attains minimum at a single point, and hence $\bar{y} = y^0$. We have thus proved that for $\alpha \to 0$ the family $\{\tilde{y}_\alpha\}$ has a single limiting point:

$$\lim_{\alpha \to 0} \tilde{y}_\alpha = y^0. \quad \triangle$$

7.4.2. Remark. Generally, it is impossible to derive a priori estimates for the rate of convergence of y_α to y^0. Simple examples indicate that the rate at which y_α tends to y^0 for $\alpha \to 0$ can be very low. For instance, in the one-dimensional problem

$$\min_{x \in E_1} x^{2k}, \quad k \geqslant 1,$$

taking $x^0 \neq 0$ and $\Omega(x) = (x - x^0)^2$ we see that y_α tends to $y^0 = 0$ for $\alpha \to 0$ at a rate of $O\left(\alpha^{\frac{1}{2k-1}}\right)$.

In the end of this chapter we shall see that this fact can markedly affect the convergence of the regularization procedure in problems where the feasible region also undergoes perturbation.

7.4.3. Theorem. *If the set X is convex and closed, the function $\varphi(x)$ is convex and continuous, the functions $\varphi_\varepsilon(x)$ are continuous,*

$$\varphi_\varepsilon(x) \in P_\varepsilon(\varphi) = \{\varphi_\varepsilon(x): \ |\varphi(x) - \varphi_\varepsilon(x)| \leqslant \varepsilon, \quad x \in X\},$$

and the set Y is nonempty, then the problem of minimizing the function $N_\alpha(x)$ on the set X is correct for any fixed $\alpha > 0$.

Proof. Denote

$$N_{\alpha\varepsilon}(x) = \varphi_\varepsilon(x) + \alpha\Omega(x).$$

Since $\varphi_\varepsilon(x) \geqslant \varphi(x) - \varepsilon \geqslant \varphi(y) - \varepsilon = \text{const}$ for all $x \in X$, the closed set $\{x \in X: N_{\alpha\varepsilon}(x) \leqslant \text{const}\}$ is bounded because $N_{\alpha\varepsilon}(x) \geqslant \text{const} + \alpha\Omega(x) \to +\infty$ for $\|x\| \to \infty$ $(x \in X)$, and, consequently, there exists

$$y_{\alpha\varepsilon} = \arg\min\{N_{\alpha\varepsilon}(x): \quad x \in X\}.$$

It can easily be seen that the functions $N_\alpha(x)$ and $N_{\alpha\varepsilon}(x)$ satisfy the conditions of Theorem 7.2.3, whence it follows that the problem under consideration is correct. △

7.5. Regularization Method (the General Case)

7.5.1. In the foregoing section we considered the regularization method for problems with exact constraints when only the objective function underwent perturbation. Below we consider problems of the type 7.1.4, i.e. the general case where only ε-information about the problem 7.1.2 is known, namely the functions $\varphi_\varepsilon(x)$ and $f_\varepsilon(x)$ belonging to the set $P_\varepsilon(\varphi, f)$ are given. To find approximate values of the normal solution to the problem 7.1.2 the following technique is used.

7.5.2. Regularization. The parametric function

$$\Psi_{\alpha\beta\varepsilon}(x) = \beta[\varphi_\varepsilon(x) + \alpha\Omega(x)] + \psi_\varepsilon(x),$$

where $\beta > 0$ and

$$\psi_\varepsilon(x) = \sum_{i=1}^{m} |\min\{f_{\varepsilon i}(x), 0\}|^p, \quad p > 0,$$

is introduced and

$$\mathbf{y}_{\alpha\beta\varepsilon} = \arg\min \{\Psi_{\alpha\beta\varepsilon}(\mathbf{x}): \mathbf{x} \in \Gamma\}$$

is calculated.

We stress that $\Psi_{\alpha\beta\varepsilon}(\mathbf{x})$ is a penalty function for the problem of minimizing the regularizing function

$$N_{\alpha\varepsilon}(\mathbf{x}) = \varphi_\varepsilon(\mathbf{x}) + \alpha\Omega(\mathbf{x})$$

on the set X_ε.

7.6. Convergence of the Regularization Method (the General Case)

7.6.1. Assumptions:

(a) the set Γ is convex and closed,

(b) the function $\varphi(\mathbf{x})$ is convex and continuous on Γ,

(c) the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are concave and continuous on Γ,

(d) the continuous functions $\varphi_\varepsilon(\mathbf{x})$ and $\mathbf{f}_\varepsilon(\mathbf{x})$ on the set Γ belong to the set $P_\varepsilon(\varphi, \mathbf{f})$,

(e) $Y = \operatorname{Arg\,min}\{\varphi(\mathbf{x}): \mathbf{x} \in X\} \neq \varnothing$,

(f)

$$\psi(\mathbf{x}) = \sum_{i=1}^{m} |\min\{f_i(\mathbf{x}), 0\}|,$$

$$\psi_\varepsilon(\mathbf{x}) = \sum_{i=1}^{m} |\min\{f_{\varepsilon i}(\mathbf{x}), 0\}|.$$

7.6.2. Proposition 1. *For any* $\mathbf{x} \in \Gamma$ *the inequality*

$$|\psi(\mathbf{x}) - \psi_\varepsilon(\mathbf{x})| \leqslant \varepsilon m$$

is fulfilled.

Proof. It is proved quite simply that

$$\big|\,|\min\{f_i(\mathbf{x}), 0\}| - |\min\{f_{\varepsilon i}(\mathbf{x}), 0\}|\,\big| \leqslant |f_i(\mathbf{x}) - f_{\varepsilon i}(\mathbf{x})|$$

(to this end it suffices to consider the cases when $f_i(\mathbf{x})$ and $f_{\varepsilon i}(\mathbf{x})$ are of one sign and of opposite signs). Further,

$$|\psi(\mathbf{x}) - \psi_\varepsilon(\mathbf{x})| \leqslant \sum_{i=1}^{m} \Big| |\min\{f_i(\mathbf{x}), 0\}| - |\min\{f_{\varepsilon i}(\mathbf{x}), 0\}| \Big| \leqslant \sum_{i=1}^{m} |f_i(\mathbf{x}) - f_{\varepsilon i}(\mathbf{x})| \leqslant \varepsilon m. \quad \triangle$$

7.6.3. Proposition 2. *For any* $\mathbf{x} \in \Gamma$, $0 \leqslant \beta \leqslant \beta_0 < \infty$, $0 \leqslant \varepsilon \leqslant \varepsilon_0 < \infty$, *and any fixed* $\alpha > 0$ *there exists*

$$\Psi^* = \inf_{x \in \Gamma} \Psi_{\alpha\beta\varepsilon}(\mathbf{x}) > -\infty.$$

Proof. Since the function

$$N_\alpha(\mathbf{x}) = \varphi(\mathbf{x}) + \alpha\Omega(\mathbf{x}), \quad \alpha > 0,$$

is strongly convex on Γ, there is

$$N^{**} = \min_{\mathbf{x} \in \Gamma} N_\alpha(\mathbf{x}) > -\infty.$$

Taking the inequality (see the condition (d))

$$\varphi_\varepsilon(\mathbf{x}) \geqslant \varphi(\mathbf{x}) - \varepsilon$$

into account we obtain

$$\Psi_{\alpha\beta\varepsilon}(\mathbf{x}) \geqslant \beta[\varphi(\mathbf{x}) - \varepsilon] + \beta\alpha\Omega(\mathbf{x}) = \beta N_\alpha(\mathbf{x}) - \beta\varepsilon \geqslant \beta N^{**} - \beta\varepsilon > -\infty. \quad \triangle$$

7.6.4. Assume that $\tilde{\mathbf{y}}_{\alpha\beta\varepsilon} \in \Gamma$ satisfies the condition

$$\Psi_{\alpha\beta\varepsilon}(\tilde{\mathbf{y}}_{\alpha\beta\varepsilon}) \leqslant \Psi^* + \eta(\varepsilon), \quad \eta(\varepsilon) > 0.$$

As before, we denote

$$\mathbf{y}^0 = \arg\min\{\Omega(\mathbf{y}): \ \mathbf{y} \in Y\},$$

$$\mathbf{y}_\alpha = \arg\min\{N_\alpha(\mathbf{x}): \ \mathbf{x} \in X\}, \quad N_\alpha(\mathbf{y}_\alpha) = N^*.$$

7.6.5. Theorem. *If the assumptions* 7.6.1 *hold and if*

$$\lim_{\varepsilon \to 0} \beta(\varepsilon) = 0, \quad \lim_{\varepsilon \to 0} \frac{\varepsilon}{\beta(\varepsilon)} = 0, \quad \lim_{\varepsilon \to 0} \frac{\eta(\varepsilon)}{\beta(\varepsilon)} = 0,$$

then

$$\lim_{\alpha \to 0} \lim_{\varepsilon \to 0} \tilde{\mathbf{y}}_{\alpha\beta\varepsilon} = \mathbf{y}^0.$$

Proof. The conditions

$$\psi_\varepsilon(\mathbf{x}) \geqslant 0,$$

$$\varphi(\mathbf{x}) \leqslant \varphi_\varepsilon(\mathbf{x}) + \varepsilon,$$

$$\varphi_\varepsilon(\mathbf{x}) \leqslant \varphi(\mathbf{x}) + \varepsilon$$

holding for all $\mathbf{x} \in \Gamma$ and also the inequality

$$\Psi_{\alpha\beta\varepsilon}(\tilde{\mathbf{y}}_{\alpha\beta\varepsilon}) \leqslant \Psi^* + \eta(\varepsilon)$$

imply

$$N^{**} \leqslant N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) \leqslant N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) + \frac{1}{\beta}\psi_\varepsilon(\tilde{y}_{\alpha\beta\varepsilon})$$

$$= \varphi(\tilde{y}_{\alpha\beta\varepsilon}) + \alpha\Omega(\tilde{y}_{\alpha\beta\varepsilon}) + \frac{1}{\beta}\psi_\varepsilon(\tilde{y}_{\alpha\beta\varepsilon})$$

$$\leqslant \varphi_\varepsilon(\tilde{y}_{\alpha\beta\varepsilon}) + \alpha\Omega(\tilde{y}_{\alpha\beta\varepsilon}) + \frac{1}{\beta}\psi_\varepsilon(\tilde{y}_{\alpha\beta\varepsilon}) + \varepsilon$$

$$= \frac{1}{\beta}\Psi_{\alpha\beta\varepsilon}(\tilde{y}_{\alpha\beta\varepsilon}) + \varepsilon \leqslant \frac{1}{\beta}(\Psi^* + \eta(\varepsilon)) + \varepsilon$$

$$\leqslant \frac{1}{\beta}\Psi_{\alpha\beta\varepsilon}(y_\alpha) + \frac{\eta(\varepsilon)}{\beta} + \varepsilon$$

$$= \varphi_\varepsilon(y_\alpha) + \alpha\Omega(y_\alpha) + \frac{1}{\beta}\psi_\varepsilon(y_\alpha) + \frac{\eta(\varepsilon)}{\beta} + \varepsilon.$$

Since $y_\alpha \in X$, it follows from 7.6.2 that

$$\psi_\varepsilon(y_\alpha) \leqslant \varepsilon m,$$

and therefore, with regard to $\varphi_\varepsilon(y_\alpha) \leqslant \varphi(y_\alpha) + \varepsilon$, we find

$$N^{**} \leqslant N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) \leqslant \varphi(y_\alpha) + \alpha\Omega(y_\alpha) + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon$$

$$= N^* + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon.$$

Since $\varepsilon m/\beta \leqslant \text{const}$ and $\eta(\varepsilon)/\beta \leqslant \text{const}$, the expression

$$N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) = \varphi(\tilde{y}_{\alpha\beta\varepsilon}) + \alpha\Omega(\tilde{y}_{\alpha\beta\varepsilon})$$

is bounded, and, since the function $N_\alpha(x)$ is strongly convex on Γ for any $\alpha > 0$, the set $\{\tilde{y}_{\alpha\beta\varepsilon}\}$ is a compactum. Let $\tilde{y}$ be a limiting point:

$$\tilde{y} = \lim_{k\to\infty} \tilde{y}_{\alpha\beta_k\varepsilon_k}.$$

From the inequality

$$N^{**} \leqslant N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) \leqslant N_\alpha(\tilde{y}_{\alpha\beta\varepsilon}) + \frac{1}{\beta}\psi_\varepsilon(\tilde{y}_{\alpha\beta\varepsilon})$$

$$\leqslant N^* + \frac{\varepsilon m}{\beta} + \frac{\eta(\varepsilon)}{\beta} + 2\varepsilon \quad (7.7)$$

it follows that

$$\lim_{k\to\infty} \psi_{\varepsilon_k}(\tilde{y}_{\alpha\beta_k\varepsilon_k}) = 0.$$

From 7.6.2 we derive

$$\psi(\tilde{\mathbf{y}}) = \lim_{k\to\infty} (\tilde{\mathbf{y}}_{\alpha\beta_k\varepsilon_k}) \leqslant \lim_{k\to\infty} [\psi_{\varepsilon_k} (\tilde{\mathbf{y}}_{\alpha\beta_k\varepsilon_k}) + \varepsilon_k m] = 0,$$

and therefore $\tilde{\mathbf{y}} \in X$. Passing to the limit in (7.7) for $\varepsilon = \varepsilon_k \to 0$ we obtain

$$N_\alpha(\tilde{\mathbf{y}}) \leqslant N^* = \min_{\mathbf{x}\in X} N_\alpha(\mathbf{x}) = N_\alpha(\mathbf{y}_\alpha),$$

and hence

$$N_\alpha(\tilde{\mathbf{y}}) = N_\alpha(\mathbf{y}_\alpha).$$

The uniqueness of the point of minimum $\mathbf{y}_\alpha$ of the strongly convex function $N_\alpha(\mathbf{x})$ on the convex set X implies $\tilde{\mathbf{y}} = \mathbf{y}_\alpha$. Thus, for $\varepsilon \to 0$ any limiting point of the compactum $\{\tilde{\mathbf{y}}_{\alpha\beta\varepsilon}\}$ coincides with $\tilde{\mathbf{y}}$, and, consequently,

$$\lim_{\varepsilon\to 0} \tilde{\mathbf{y}}_{\alpha\beta\varepsilon} = \mathbf{y}_\alpha.$$

It follows from Theorem 7.4.1 that

$$\lim_{\alpha\to 0} \mathbf{y}_\alpha = \mathbf{y}^0;$$

and therefore

$$\lim_{\alpha\to 0} \lim_{\varepsilon\to 0} \tilde{\mathbf{y}}_{\alpha\beta\varepsilon} = \mathbf{y}^0. \quad \triangle$$

7.6.6. Remark. The foregoing theorem remains true when the smooth penalty

$$\psi(\mathbf{x}) = \sum_{i=1}^{m} [\min\{f_i(\mathbf{x}), 0\}]^2,$$

$$\psi_\varepsilon(\mathbf{x}) = \sum_{i=1}^{m} [\min\{f_{\varepsilon i}(\mathbf{x}), 0\}]^2$$

is taken, and in this case the condition $\lim_{\varepsilon\to 0} \frac{\varepsilon}{\beta(\varepsilon)} = 0$ is replaced by $\lim_{\varepsilon\to 0} \frac{\varepsilon^2}{\beta(\varepsilon)} = 0$. The proof is carried out according to the same scheme. The distinction lies in that Proposition 7.6.2 is changed, namely, for any $\mathbf{x} \in \Gamma$ we have $|\psi(\mathbf{x}) - \psi_\varepsilon(\mathbf{x})| \leqslant m\varepsilon^2 + \varepsilon\sqrt{m}\,(1 + \psi_\varepsilon(\mathbf{x}))$.

7.6.7. Since in real situations it often happens that the value of the parameter ε cannot be improved (diminished), it also characterizes the maximum accuracy with which it makes sense to calculate the point of minimum (see Sec. 7.6.4 and the condition $\frac{\eta(\varepsilon)}{\beta(\varepsilon)} \to 0$ for $\varepsilon \to 0$). The parameter ε also restricts the choice of the parameter $\beta = \beta(\varepsilon)$. The condition

$$\frac{\varepsilon}{\beta(\varepsilon)} \to 0 \quad \text{for} \quad \varepsilon \to 0$$

indicates that the choice of $\beta(\varepsilon)$ depends on the given value of ε.

The regularization procedure encounters a number of difficulties. We consider only one of them, namely the selection of the initial values α_0 and β_0 of the regularization parameters. We shall assume that ε is sufficiently small. Since the regularization process may converge very slowly (see Remark 7.4.2), the choice of a comparatively large value of α_0 may be extremely disadvantageous from the standpoint of the total computing time required to obtain a sufficiently accurate approximation to the normal solution.

On the other hand, as a rule, to determine the point $y_{\alpha_0\beta_0\varepsilon}$ use is made of relaxational minimization methods whose rate of convergence is strongly affected by the value of the parameter of strong convexity of the function $\Psi_{\alpha_0\beta_0\varepsilon}(x)$ under minimization, i.e. the value of the product $\alpha_0\beta_0$. As is known, when $\alpha\beta$ decreases, the rate of convergence of relaxation processes is reduced. Therefore the choice of small values of α_0 and β_0 may also increase the total computing time required to obtain the desired approximation to the normal solution and, besides, it may prove unjustifiable where compatibility with the value of ε is concerned.

In conclusion we note that in this chapter we considered questions of stability in mathematical programming problems. However, of importance also is the question of stability of numerical minimization methods relative to possible errors appearing in the computation process. The reader can find the analysis of this factor in the subsequent chapters of the present book.

Chapter 8

ONE-DIMENSIONAL MINIMIZATION METHODS

8.1. On One-Dimensional Minimization Problems

8.1.1. Minimization of a function of one variable is, as a rule, a necessary element in many minimization methods for multi-dimensional functions.

At first glance, it may seem that the problem of minimizing a function of one variable is quite simple. Perhaps such an opinion is formed when one begins to study fundamentals of calculus. Indeed, if a function $\varphi(x)$ to be minimized on a closed interval $[a, b]$ is differentiable, then it suffices to determine the zeros of the derivative, add the end points of the interval to them, find the local minima among these points, and, finally, take the point where the absolute minimum is attained.

However, for a wide class of functions this problem is not at all so simple. First, in real situations the problem of solving the equation $\varphi'(x) = 0$ may turn out to be rather complicated. On the other hand, in practical problems it is often unknown whether the function $\varphi(x)$ is differentiable or not.

Therefore of major importance are minimization methods not requiring the calculation of derivatives.

8.1.2. If it is only known that the function $\varphi(x)$ under minimization is continuous on the interval $[a, b]$, then to determine the point of minimum it is necessary to study the behavior of the function at all the points of the interval. Since we are interested in an approximate calculation of the point of minimum, the behavior of the function is investigated at a finite number of points, and one-dimensional minimization methods differ in the ways these points are chosen. The majority of deterministic methods consists in constructing a sequence of intervals $[a_n, b_n]$ contracting to the point $x^* = \arg\min \{\varphi(x): x \in [a, b]\}$. However, there always exists a continuous

function $\varphi(x)$ such that for any finite index $n > 0$ the interval $[a_n, b_n]$ constructed by means of any deterministic method does not contain the point x^* even when an interval $[a_0, b_0]$ to which x^* belongs is known. Therefore one can guarantee that the point x^* belongs to the interval $[a_n, b_n]$ only for certain definite classes of functions. In what follows we confine ourselves to the class of strictly unimodal functions and sometimes even to a narrower class.

8.1.3. Strictly unimodal functions. A continuous function $\varphi(x)$ will be called *strictly unimodal* if it has a single point of minimum x^* and

$$\varphi(x_1) > \varphi(x_2) \text{ for any } x_1 < x_2 \leqslant x^*,$$

$$\varphi(x_1) < \varphi(x_2) \text{ for any } x^* \leqslant x_1 < x_2.$$

Thus, with increasing x, the function $\varphi(x)$ decreases monotonically on the left of the point of minimum and increases monotonically on the right of it.

8.1.4. Before beginning the minimization procedure one should make sure that the objective function belongs to a "suitable" class, i.e. a class for which the convergence of the minimization process is guaranteed. This does not at all mean that the attempts to apply deterministic methods to minimize functions which are not known to belong to a "suitable" class or have no such a class are doomed to failure. However, in many cases there exists additional information making it possible to indicate the needed class. This information is often related to the "physical" meaning of the actual problem for which the mathematical model under study is constructed. We note that, generally, the assumption that a function is strictly unimodal in a neighborhood of the point x^* is quite natural, and therefore the search for information about this neighborhood is an important preliminary stage in the minimization procedure.

In what follows we shall use the notion of an appropriate triple of numbers.

8.1.5. An appropriate triple of numbers. We shall denote $\varphi(x_i) = \varphi_i$ $(i = 1, 2, \ldots)$.

A triple of numbers $x_1 < x_2 < x_3$ is said to be *appropriate* if

$$\varphi_2 \leqslant \min\{\varphi_1, \varphi_3\}$$

and

$$\varphi_2 < \max \{\varphi_1, \varphi_3\}.$$

The latter condition implies that the points (x_1, φ_1), (x_2, φ_2), and (x_3, φ_3) do not lie in a straight line parallel to the axis of abscissas (naturally, for a strictly unimodal function this parallelism is impossible). Thus, an appropriate triple of numbers is such that the closed interval $[x_1, x_3]$ is sure to contain x^*, the point of minimum of the strictly unimodal function, as its interior point.

8.1.6. On the choice of an appropriate triple of numbers. In deterministic minimization methods each calculation cycle usually results in the appearance of points (x_1, φ_1), (x_2, φ_2), (x_3, φ_3), and (x_4, φ_4) such that

$$x_1 < x_2 < x_3 < x_4,$$

$$\varphi_2 \leqslant \min \{\varphi_1, \varphi_4\}, \quad \varphi_3 \leqslant \min \{\varphi_1, \varphi_4\}.$$

There arises the problem of selecting an appropriate triple. In view of the strict unimodality of the minimized function, it is obvious that for $\varphi_2 < \varphi_3$ the triple x_1, x_2, x_3 is appropriate and for $\varphi_2 > \varphi_3$ an appropriate triple is formed by the numbers x_2, x_3, x_4. In case $\varphi_2 = \varphi_3$, we have $x^* \in [x_2, x_3]$, and any of the triples x_1, x_2, x_3 and x_2, x_3, x_4 is appropriate. It is advisable to choose among them the triple to which the shortest of the two intervals $[x_1, x_3]$ and $[x_2, x_4]$ corresponds. However, one can also apply an alternative technique, e.g. to use the division of an interval in two in order to calculate a point x_5 belonging to the interval $[x_2, x_3]$. The triple x_2, x_5, x_3 is sure to be appropriate.

Since the calculations in the numerical realization of the method are limited to a definite accuracy, one can only assert that the values φ_2 and φ_3 are equal to an ε-accuracy, i.e. $|\varphi_2 - \varphi_3| \leqslant \varepsilon$. In this situation each of the triples x_1, x_2, x_3 and x_2, x_3, x_4 may turn to be appropriate. Then to select the triple in a suitable manner it is advisable to calculate the value φ_5 at the point $x_5 = (x_2 + x_3)/2$. If φ_5 coincides with φ_2 and φ_3 to an accuracy of ε, one stops the search and puts $x^* \simeq x_5$. If otherwise, x_2, x_5, x_3 is an appropriate triple.

8.2. Search for an Interval Containing the Point of Minimum

8.2.1. The search consists in performing steps with increasing lengths until the point of minimum of the function $\varphi(x)$ has been passed.

8.2.2. *Step 1* (nonstandard). A point x_0 is chosen and the direction of decrease of the function $\varphi(x)$ is determined. To this end we take a number h and calculate $\varphi(x_0 + h)$. If $\varphi(x_0 + h) < \varphi(x_0)$, we put $x_1 = x_0 + h$ and pass to step 2 for $k = 1$. If $\varphi(x_0 + h) \geqslant \varphi(x_0)$, we put $h = -h$ and calculate $\varphi(x_0 + h)$. If $\varphi(x_0 + h) < \varphi(x_0)$, we put $x_1 = x_0 + h$ and pass to step 2 for $k = 1$. If $\varphi(x_0 + h) \geqslant \varphi(x_0)$, we put $h = h/2$ and repeat the procedure of the foregoing step. Step 1 results in a number h and points x_0, $x_1 = x_0 + h$ such that $\varphi(x_1) < \varphi(x_0)$.

Step 2. We duplicate h and calculate $x_{k+1} = x_k + h$.

Step 3. φ_{k+1} is calculated. If $\varphi_{k+1} < \varphi_k$, one puts $k = k + 1$ and passes to step 2. If $\varphi_{k+1} \geqslant \varphi_k$, the search is stopped and $[x_{k-1}, x_{k+1}]$ is taken as an interval containing the point of minimum.

8.3. Fibonacci and Golden Section Methods

8.3.1. Computation procedures often encounter essential difficulties related to the calculation of the values of a function $\varphi(x)$ when, for instance, the function is found in an experiment or is represented analytically and the corresponding calculation formula is very complicated. Among the methods where, in a sense, the best accuracy is attained when the number of the calculated values of $\varphi(x)$ is limited we mention the Fibonacci and golden section methods.

The methods consist in constructing sequences of closed intervals $[a_k, b_k]$ contracting to the point x^* of minimum of the function $\varphi(x)$ on the original interval $[a_0, b_0]$.

8.3.2. The scheme of the methods. The two methods are constructed according to a unified scheme and differ in the choice of a parameter λ.

Step 1 (nonstandard). An interval $[a_0, b_0]$ containing the point x^* of minimum of the function $\varphi(x)$ is chosen;

an accuracy ε with which the point x^* should be determined is set; $\Delta_0 = b_0 - a_0$, $\Delta_1 = \lambda\Delta_0$, $1/2 < \lambda < 1$, $\Delta_2 = \Delta_0 - \Delta_1$, $y_0 = a_0 + \Delta_2$, $z_0 = b_0 - \Delta_2$, $\varphi(y_0)$, and $\varphi(z_0)$ are calculated, and then one puts $k = 1$.

Step 2. (a) If $\varphi(y_{k-1}) \leqslant \varphi(z_{k-1})$, one puts $a_k = a_{k-1}$, $b_k = z_{k-1}$, $z_k = y_{k-1}$, calculates $\Delta_{k+2} = \Delta_k - \Delta_{k+1}$, $y_k = a_k + \Delta_{k+2}$, $\varphi(y_k)$, and passes to step 3.

(b) If $\varphi(y_{k-1}) > \varphi(z_{k-1})$, one puts $a_k = y_{k-1}$, $b_k = b_{k-1}$, $y_k = z_{k-1}$, calculates $\Delta_{k+2} = \Delta_k - \Delta_{k+1}$, $z_k = b_k - \Delta_{k+2}$, $\varphi(z_k)$, and passes to step 3.

Step 3. If $\Delta_k \leqslant \varepsilon$, one puts $x^* \simeq x_k = \arg\min\{\varphi(y_k), \varphi(z_k)\}$. If $\Delta_k > \varepsilon$, one puts $k = k + 1$ and passes to step 2.

8.3.3. It should be noted that at each step, except the first one, only one calculation of the value of the function $\varphi(x)$ is performed.

8.3.4. Analysis of the method. We shall find out what are the conditions on the number λ under which each closed interval $[a_k, b_k]$ contains the point x^*, $[a_k, b_k] \subset [a_{k-1}, b_{k-1}]$ and, finally, these intervals contract to the point x^* with increasing k.

The conditions

$$\Delta_0,\ \Delta_1 = \lambda\Delta_0,\ \Delta_{k+2} = \Delta_k - \Delta_{k+1},\ k = 0, 1, \ldots, \quad (8.1)$$

and the above scheme imply that

8.3.5. *If* $\Delta_k > 0$ *for all* $k = 0, 1, \ldots,$ *then*

$$\Delta_k = b_k - a_k.$$

Indeed, for $k = 0$ this follows from the definition of the number Δ_0. We state the induction hypothesis that $\Delta_{k-1} = b_{k-1} - a_{k-1}$. By construction, $\Delta_{k+1} = b_{k-1} - z_{k-1} = y_{k-1} - a_{k-1}$. However, $\Delta_k = \Delta_{k-1} - \Delta_{k+1}$, and, consequently,

$$\Delta_k = (b_{k-1} - a_{k-1}) - (b_{k-1} - z_{k-1}) = z_{k-1} - a_{k-1},$$

i.e.

$$\Delta_k = (b_{k-1} - a_{k-1}) - (y_{k-1} - a_{k-1}) = b_{k-1} - y_{k-1}.$$

If substep (a) takes place in step 2, then $b_k = z_{k-1}$ and $a_k = a_{k-1}$, therefore $\Delta_k = b_k - a_k$. In case substep (b) takes place, we have $b_k = b_{k-1}$, $a_k = y_{k-1}$, and hence $\Delta_k = b_k - a_k$. $\triangle$

8.3.6. If $\Delta_m = 0$, then $\Delta_{m-1} = \Delta_{m-2}$, and, consequently, $y_{m-3} = z_{m-3}$, and therefore the process of constructing the intervals $[a_k, b_k]$ is stopped at $k = m - 2$.

8.3.7. By induction, (8.1) readily implies

$$\Delta_k = \Delta_k(\lambda) = (-1)^{k-1}(F_k\lambda - F_{k-1})\Delta_0, \quad k = 2, 3, \ldots, \tag{8.2}$$

$$F_1 = F_2 = 1, \quad F_{k+1} = F_k + F_{k-1}, \quad k = 2, 3, \ldots. \tag{8.3}$$

8.3.8. The numbers F_k satisfying (8.3) are called *Fibonacci numbers*. We present the first few of them: $F_3 = 2$, $F_4 = 3$, $F_5 = 5$, $F_6 = 8$, $F_7 = 13$, $F_8 = 21$, $F_9 = 34$, $F_{10} = 55$, $F_{11} = 89$, $F_{12} = 144$, etc. For our further consideration we need some properties of the Fibonacci numbers.

8.3.9. *Binet's formula*

$$F_k = \frac{1}{\sqrt{5}}\left[\left(\frac{1+\sqrt{5}}{2}\right)^k - \left(\frac{1-\sqrt{5}}{2}\right)^k\right], \quad k = 1, 2, \ldots$$

is needed here. Its validity can be checked by substituting it into (8.3) taking into consideration that the numbers F_k are determined uniquely for each index k.

8.3.10. The relation

$$F_{k+1}^2 = F_kF_{k+2} + (-1)^k, \quad k = 1, 2, \ldots \tag{8.4}$$

can easily be proved by induction on k.

8.3.11. Let $\lambda_{k+1} = F_k/F_{k+1}$ $(k = 1, 2, \ldots)$. Using Binet's formula we find:

$$\lambda_{k+1} = \frac{\sqrt{5}-1}{2}\,\frac{1+(-1)^k\left(\frac{3-\sqrt{5}}{2}\right)^k}{1+(-1)^{k+1}\left(\frac{3-\sqrt{5}}{2}\right)^{k+1}},$$

whence it follows that $\lim\limits_{k\to\infty} \lambda_k = \dfrac{\sqrt{5}-1}{2}$.

8.3.12. From the definition of the number λ_k and formula (8.4) we derive

$$\lambda_k - \lambda_{k+1} = \frac{F_{k-1}F_{k+1} - F_k^2}{F_kF_{k+1}} = \frac{(-1)^k}{F_kF_{k+1}}, \quad k = 2, 3, \ldots$$

8.3.13. Therefore

$$\lambda_{2m+2} < \lambda_{2m}, \quad m = 1, 2, \ldots,$$
$$\text{and} \quad \lambda_{2m+1} > \lambda_{2m-1}, \; m = 1, 2, \ldots$$

8.3.14. *Let* $k \leqslant n$. *Then* $[a_k, b_k] \subset [a_{k-1}, b_{k-1}]$ *if* $\lambda_n = F_{n-1}/F_n$, *and*

$$\begin{aligned} \lambda_n < \lambda < \lambda_{n+1} \quad & \text{if } n \text{ is odd}, \\ \lambda_{n+1} < \lambda < \lambda_n \quad & \text{if } n \text{ is even}. \end{aligned} \tag{8.5}$$

Indeed, the scheme of the method and 8.3.5 imply that the condition $[a_k, b_k] \subset [a_{k-1}, b_{k-1}]$ is equivalent to

$$0 < \Delta_k(\lambda) < \Delta_{k-1}(\lambda). \tag{8.6}$$

If k is odd, then from (8.2) we conclude that for $\Delta_k(\lambda) > 0$ it is necessary that the inequality $\lambda > \lambda_k$ hold, and for $\Delta_k(\lambda) < \Delta_{k-1}(\lambda)$ there must be $\lambda > \lambda_{k+1}$. For even k we similarly obtain $\lambda_{k+1} < \lambda < \lambda_k$. Thus, the following inequalities guarantee the validity of the condition (8.6):

$$\begin{aligned} \lambda_k < \lambda < \lambda_{k+1} \quad & \text{if } k \text{ is odd}, \quad k \leqslant n, \\ \lambda_{k+1} < \lambda < \lambda_k \quad & \text{if } k \text{ is even}, \quad k \leqslant n. \end{aligned} \tag{8.7}$$

It follows from 8.3.13 and 8.3.11 that for $m \to \infty$ we have $\lambda_{2m+1} \uparrow \bar{\lambda}$, $\lambda_{2m} \downarrow \bar{\lambda}$, $\bar{\lambda} = \dfrac{\sqrt{5}-1}{2}$, therefore, if the conditions (8.5) hold, then the conditions (8.7) also hold, and hence (8.6) is fulfilled. As is seen from the scheme of the method, each of the intervals $[a_k, b_k]$ contains the point x^*. Δ

8.3.15. The following assertion holds: *for* $\lambda = \bar{\lambda}$ *the presented method generates an infinite sequence of nested closed intervals* $[a_k, b_k]$ *contracting to the point* x^*.

Indeed, since $\bar{\lambda}^2 = 1 - \bar{\lambda}$, by induction on n we derive from (8.1):

$$\Delta_n(\bar{\lambda}) = \bar{\lambda}^n \Delta_0 = \left(\frac{\sqrt{5}-1}{2}\right)^n \Delta_0 < 0.7^n \Delta_0.$$

Therefore, $\Delta_n(\bar{\lambda}) \to 0$ for $n \to \infty$.

Finally, since the point x^* belongs to all the intervals $[a_k, b_k]$, the concluding part of the assertion is true. Δ

8.3.16. The golden section method. For $\lambda = \bar{\lambda}$ the above method is called the *golden section method.* We elucidate this name. A point y_k is said to perform the golden section of an interval $[a_k, b_k]$ if

$$\frac{b_k - a_k}{b_k - y_k} = \frac{b_k - y_k}{y_k - a_k}.$$

It can easily be shown (check this) that for $\lambda = \bar{\lambda}$ each of the points y_k and z_k constructed according to the scheme of the method performs the golden section:

$$\frac{b_k - a_k}{b_k - y_k} = \frac{b_k - y_k}{y_k - a_k} = \frac{b_k - a_k}{z_k - a_k} = \frac{z_k - a_k}{b_k - z_k} = \frac{1 + \sqrt{5}}{2}.$$

8.3.17. We now pose the following question: for what values of the parameter λ we obtain, after a finite number of steps, the shortest interval $[a_k, b_k]$?

Proposition. *If* $\lambda = \lambda_n$, *then for any* λ *satisfying the conditions* (8.5) *the inequality*

$$0 < \Delta_{n-2}(\lambda_n) < \Delta_{n-2}(\lambda)$$

holds.

Indeed, (8.2) implies

$$\Delta_{n-2}(\lambda_n) - \Delta_{n-2}(\lambda) = (-1)^{n-3} F_{n-2}(\lambda_n - \lambda)\,\Delta_0.$$

If n is odd, then (8.5) implies that $\lambda_n < \lambda$, i.e. $\Delta_{n-2}(\lambda_n) - \Delta_{n-2}(\lambda) < 0$. Now, since, by virtue of (8.5), for even n we have $\lambda_n > \lambda$, it follows that $\Delta_{n-2}(\lambda_n) - \Delta_{n-2}(\lambda) < 0$. $\triangle$

8.3.18. There is no sense in performing more than $n - 2$ steps for $\lambda = \lambda_n$. Since $\Delta_n(\lambda_n) = 0$ and $\Delta_{n-1}(\lambda_n) = \Delta_{n-2}(\lambda_n)$, we have $y_{n-1} = z_{n-1}$, and therefore the process is stopped at the $(n - 2)$ th step.

8.3.19. The optimality of the Fibonacci method. For $\lambda = \lambda_n$ the method under study is called the *Fibonacci method.* It is optimal in the sense of the proposition in Sec. 8.3.17.

8.3.20. The sensitivity of the method to errors. The numerical realization of both methods possesses an essential shortcoming. The matter is that approximate values of the numbers $\bar{\lambda}$ and λ_n result in accumulation of errors such that it can very soon lead to violation of the condition that the intervals $[a_k, b_k]$ are nested and thus to the

divergence of the process. Indeed, let $\hat{\lambda}$ satisfy (8.5). Consider $\lambda = \hat{\lambda} + \delta$. From (8.2) we derive

$$\Delta_k(\lambda) = (-1)^{k-1}(F_k\lambda - F_{k-1})\Delta_0$$
$$= \Delta_k(\hat{\lambda}) + (-1)^{k-1}\delta F_k\Delta_0. \quad (8.8)$$

Since the inequality $0 < \Delta_k(\hat{\lambda}) < \Delta_{k-1}(\hat{\lambda})$ holds, the sequence $\{\Delta_k(\hat{\lambda})\}$ is convergent for $k \to \infty$. At the same time, $F_k \to +\infty$ and therefore $|\Delta_k(\lambda)| \to +\infty$.

It is known that Fibonacci numbers increase at a rate of a geometric progression with common ratio $\frac{1+\sqrt{5}}{2} \simeq 1.618$. As a consequence, for a fixed accuracy the method "disbalances" rather fast.

8.3.21. The choice of the number n. Given ε, we shall determine the number n necessary for guaranteeing the inequality $\Delta_{n-2} \leqslant \varepsilon$.

To this end we first prove the formulas

$$\Delta_k(\lambda_n) = \frac{F_{n-k}}{F_n}\Delta_0, \quad k = \overline{1,\ n-1}. \quad (8.9)$$

Let $k = n - l$ $(l = \overline{1,\ n-1})$. We shall show that

$$\Delta_{n-l}(\lambda_n) = \frac{F_l}{F_n}\Delta_0, \quad l = \overline{1,\ n-1}. \quad (8.10)$$

Using (8.4) we derive from (8.2):

$$\Delta_{n-l}(\lambda_n) = (-1)^{n-2}\frac{F_{n-1}^2 - F_nF_{n-2}}{F_n}\Delta_0 = \frac{1}{F_n}\Delta_0 = \frac{F_1}{F_n}\Delta_0.$$

Since $\Delta_n(\lambda_n) = 0$, (8.1) implies $\Delta_{n-2}(\lambda_n) = \Delta_{n-1}(\lambda_n) = \frac{1}{F_n}\Delta_0 = \frac{F_2}{F_n}\Delta_0$. The same formula (8.1) implies $\Delta_{n-l} = \Delta_{n-l+1} + \Delta_{n+l+2}$, whence, by induction on l, we prove (8.10) and hence (8.9). $\triangle$

Thus, to attain an ε-accuracy of the result the number n should be determined from the condition

$$\Delta_{n-2}(\lambda_n) = \frac{1}{F_n}\Delta_0 \leqslant \varepsilon.$$

8.3.22. On the compatibility of accuracy in the Fibonacci method. We now turn to a natural situation when the number λ_n is found approximately: $\lambda = \lambda_n + \delta$.

From (8.8) and (8.9) we obtain

$$\Delta_k(\lambda) = \frac{F_{n-k}}{F_n}\Delta_0 + (-1)^{k-1}\delta F_k \Delta_0. \qquad (8.11)$$

It follows that to attain the desired result the accuracy δ in the calculation of the parameter λ_n must be compatible with the preassigned accuracy ε in the determination of the length of the interval $[a_{n-2}, b_{n-2}]$. Namely, if n is chosen so that

$$\frac{1}{F_n}\Delta_0 \leqslant \frac{\varepsilon}{2},$$

then to guarantee the inequality $\Delta_{n-2}(\lambda) \leqslant \varepsilon$ there must be

$$|\delta| \leqslant \frac{\varepsilon}{2F_{n-2}\Delta_0}.$$

8.3.23. On the compatibility of accuracy in the golden section method. We consider the case when $\lambda = \bar{\lambda} + \delta$, $\bar{\lambda} = \frac{\sqrt{5}-1}{2}$. From 8.3.15 we find

$$\Delta_k(\bar{\lambda}) = \left(\frac{\sqrt{5}-1}{2}\right)^k \Delta_0. \qquad (8.12)$$

This and (8.8) imply

$$\Delta_k(\lambda) = \left(\frac{\sqrt{5}-1}{2}\right)^k \Delta_0 + (-1)^{k-1}\delta F_k \Delta_0, \qquad (8.13)$$

and for

$$\left(\frac{\sqrt{5}-1}{2}\right)^k \Delta_0 \leqslant \frac{\varepsilon}{2} \quad \text{and} \quad |\delta| \leqslant \frac{\varepsilon}{2F_k\Delta_0}$$

we obtain $\Delta_k(\lambda) \leqslant \varepsilon$.

8.3.24. Once again on the instability. It is obvious that due to the limited accuracy of computer calculations the number n is not large and the desired ε-accuracy of the result may not be attained when the suggested scheme is applied. In this sense *both methods are not stable relative to the errors in the determination of the feasible values of the parameter* λ.

8.3.25. A modification of the method. In case the parameter δ determining the accuracy in the calculation of λ_n is preassigned and the conditions of compatibility of ε and δ (see Sec. 8.3.22 and 8.3.23) are not fulfilled, the method can be modified so that the process becomes stable.

As was already seen (see Sec. 8.3.14), the conditions $0 < \Delta_{k+1}(\lambda) < \Delta_k(\lambda)$ guarantee the inclusion $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$. It follows from (8.11) that

$$\Delta_k(\lambda) - \Delta_{k+1}(\lambda) = \frac{F_{n-k-2}}{F_n}\Delta_0 + (-1)^{k-1}\delta F_{k+2}\Delta_0,$$

and for $k = n-3$ we see that $\Delta_{n-2}(\lambda) < \Delta_{n-3}(\lambda)$ provided that

$$|\delta| < \frac{1}{F_n F_{n-1}}. \tag{8.14}$$

Thus, the condition (8.14) guarantees that the relations

$$[a_{k+1}, b_{k+1}] \subset [a_k, b_k], \quad k = \overline{0, n-3},$$

are fulfilled.

We take

$$n = \arg\max_k \left\{k: |\delta| < \frac{1}{F_k F_{k-1}}\right\} \tag{8.15}$$

and perform $n-2$ steps according to the scheme of the method. After that we put $n-2=0$ and pass to step 1. Thus, after $n-2$ steps the process is renewed, i.e. we put $a_{n-2} = a_0$, $b_{n-2} = b_0$, calculate $\Delta_0 = b_0 - a_0$, $\Delta_1 = \lambda\Delta_0$, $\Delta_2 = \Delta_0 - \Delta_1$, y_0, z_0, $\varphi(y_0)$ and $\varphi(z_0)$, and pass to step 2. After each such cycle consisting of $n-2$ steps we find, taking into consideration (8.11) and (8.15),

$$\begin{aligned}\Delta_{n-2}(\lambda) &= \frac{F_2}{F_n}\Delta_0 + (-1)^{n-3}\delta F_{n-2}\Delta_0 \\ &\leqslant \frac{\Delta_0}{F_n} + |\delta| F_{n-2}\Delta_0 < \frac{\Delta_0}{F_n} + \frac{F_{n-2}}{F_n F_{n-1}}\Delta_0 \\ &= \frac{\Delta_0}{F_n}\left(1 + \frac{F_{n-2}}{F_{n-1}}\right) \leqslant \frac{2}{F_n}\Delta_0,\end{aligned}$$

and after p cycles there must be

$$\Delta_{p(n-2)}(\lambda) < \left(\frac{2}{F_n}\right)^p \Delta_0.$$

It is obvious that $\Delta_{p(n-2)}(\lambda) \to 0$ for $p \to \infty$ when $n \geqslant 3$.

We note that in each cycle $n-1$ values of the minimized function are calculated, therefore the total number of these calculations amounts to $pn - p$ whereas in the "ideal" scheme of the method 8.3.2 one performs $pn - 2p - 1$ calculations during $p(n-2)$ steps.

8.3.26. In the golden section method, for $\lambda = \bar{\lambda} + \delta$ we derive from (8.13):

$$\Delta_k(\lambda) - \Delta_{k+1}(\lambda) = \left(\frac{\sqrt{5}-1}{2}\right)^{k+2} \Delta_0 + (-1)^{k-1} \delta F_{k+2} \Delta_0,$$

and for

$$|\delta| < \frac{1}{F_{k+2}} \left(\frac{\sqrt{5}-1}{2}\right)^{k+2} \tag{8.16}$$

we have $[a_{i+1}, b_{i+1}] \subset [a_i, b_i]$ $(i = \overline{0, k})$.

We take

$$n = \arg\max \left\{k: |\delta| < \frac{1}{F_{k+2}} \left(\frac{\sqrt{5}-1}{2}\right)^{k+2}\right\} \tag{8.17}$$

and perform n steps according to the scheme of the method. Then we put $n = 0$ and pass to step 1. It is obvious that, by analogy with the foregoing section, $\Delta_{pn}(\lambda) \to 0$ for $p \to \infty$.

8.3.27. However, the accumulation of errors may result in a very slow decrease of Δ_k even when the number n is found from (8.15) and (8.17). Therefore the continuation of the process may at some moment become "unprofitable".

Let us estimate the rate of decrease of Δ_k in the ideal scheme. It follows from (8.9) that

$$\Delta_{k+1}(\lambda_n) = \lambda_{n-k} \Delta_k(\lambda_n).$$

In particular

$$\begin{aligned}
&\Delta_{n-2}(\lambda_n) = 0.5\Delta_{n-3}(\lambda_n),\\
&\Delta_{n-3}(\lambda_n) \simeq 0.67\Delta_{n-4}(\lambda_n),\\
&\Delta_{n-4}(\lambda_n) = 0.6\Delta_{n-5}(\lambda_n),\\
&\Delta_{n-5}(\lambda_n) \simeq 0.62\Delta_{n-6}(\lambda_n),\\
&\Delta_{n-6}(\lambda_n) \simeq 0.62\Delta_{n-7}(\lambda_n), \ \ldots\ .
\end{aligned}$$

Now, (8.12) implies

$$\Delta_{k+1}(\bar{\lambda}) = \frac{\sqrt{5}-1}{2} \Delta_k(\bar{\lambda}) \simeq 0.62\Delta_k(\bar{\lambda}).$$

Therefore it is advisable to take a number $0.62 < \lambda < 1$, e.g. $\alpha = 0.8$, and to continue the process as long as $\Delta_{k+1} \leqslant \alpha\Delta_k$. For $\Delta_{k+1} > \alpha\Delta_k$ the cycle should be ended and the process should be renewed, i.e. one should put $k = 0$ and pass to step 1.

8.3.28. The modified golden section algorithm.

Step 1. Select an interval $[a_0, b_0]$, set an accuracy ε in the determination of the point x^*, and take a number $\alpha \in [0.7, 0.9]$.

Step 2. Calculate $\Delta_0 = b_0 - a_0$, $\Delta_1 = \frac{\sqrt{5}-1}{2}\Delta_0$, $\Delta_2 = \Delta_0 - \Delta_1$, $y_0 = a_0 + \Delta_2$, $z_0 = b_0 - \Delta_2$, $\varphi(y_0)$, $\varphi(z_0)$ and put $k = 1$.

Step 3. (a) If $\varphi(y_{k-1}) \leqslant \varphi(z_{k-1})$, put $a_k = a_{k-1}$, $b_k = z_{k-1}$, $z_k = y_{k-1}$ and calculate $\Delta_{k+2} = \Delta_k - \Delta_{k+1}$, $y_k = a_k + \Delta_{k+2}$. For $y_k < z_k$ calculate $\varphi(y_k)$ and pass to step 4. For $y_k \geqslant z_k$ put $a_0 = a_k$, $b_0 = b_k$ and pass to step 2.

(b) If $\varphi(y_{k-1}) > \varphi(z_{k-1})$, put $a_k = y_{k-1}$, $b_k = b_{k-1}$, $y_k = z_{k-1}$ and calculate $\Delta_{k+2} = \Delta_k - \Delta_{k+1}$, $z_k = b_k - \Delta_{k+2}$. For $z_k > y_k$ calculate $\varphi(z_k)$ and pass to step 4. For $z_k \leqslant y_k$ put $a_0 = a_k$, $b_0 = b_k$ and pass to step 2.

Step 4. (a) If $\Delta_k \leqslant \varepsilon$, stop the search and put

$$x^* \simeq x_k = \arg\min \{\varphi(y_k), \varphi(z_k)\}.$$

(b) If $\varepsilon < \Delta_k \leqslant \alpha\Delta_{k-1}$, pass to step 3 for $k = k + 1$.

(c) If $\Delta_k > \varepsilon$ and $\Delta_k > \alpha\Delta_{k-1}$, put $a_0 = a_k$, $b_0 = b_k$ and pass to step 2.

8.4. Parabolic Method

8.4.1. The idea of the method is quite clear: if a function $\varphi(x)$ is approximated sufficiently accurately by a polynomial of the second degree on an interval $[a, b]$ with an interior point $x^* = \arg\min \{\varphi(x): x \in [a, b]\}$, then it is advisable to take the point of minimum of the polynomial as an approximate value of x^*. Naturally, the suggested idea must be confirmed by a reasonable argument. When the function is smooth and strictly unimodal, one can assume that, generally, it is close to a parabola in the neighborhood of the point x^*. Therefore, it is expedient to apply the parabolic method after a sufficiently short interval containing x^* is found, e.g. after $[a_k, b_k]$ and $x_k = \arg\min \{\varphi(y_k), \varphi(z_k)\}$ are determined as a result of k steps performed according to the golden section method.

8.4.2. Let an appropriate triple of numbers x_1, x_2, x_3 (in the sense of the Definition 8.1.5) be found. The polynomial of the second degree whose graph passes through the points (x_1, φ_1), (x_2, φ_2), (x_3, φ_3) attains minimum at the point

$$\tilde{x} = x_2 - \frac{1}{2} \frac{(x_2 - x_1)^2 (\varphi_2 - \varphi_3) - (x_3 - x_2)^2 (\varphi_2 - \varphi_1)}{(x_2 - x_1)(\varphi_2 - \varphi_3) - (x_3 - x_2)(\varphi_2 - \varphi_1)}. \quad (8.18)$$

This can be checked quite simply by constructing an interpolation polynomial from the above-mentioned three points and finding its minimum. It can also be readily shown that $\tilde{x} \in [x_1, x_3]$ by verifying the inequality

$$|\tilde{x} - x_2| \leqslant \frac{1}{2} \max\{x_2 - x_1, x_3 - x_2\}.$$

8.4.3. The scheme of the method. It is assumed that an appropriate triple of numbers x_1, x_2, x_3 and the corresponding values φ_1, φ_2, φ_3 are known and an accuracy ε of the determination of an interval containing the point x^* is set.

Step 1. Calculate $\tilde{x}$ by formula (8.18).

Step 2. If $|\tilde{x} - x_2| \leqslant \varepsilon$, stop the search and put $x^* \sim \tilde{x}$. If $|\tilde{x} - x_2| > \varepsilon$, calculate $\varphi(\tilde{x})$.

Step 3. Using the method of Sec. 8.1.6, select an appropriate triple among the numbers x_1, x_2, x_3, and $\tilde{x}$ and pass to step 1.

8.5. Cubic Approximation Method

8.5.1. This method is used to minimize a differentiable function $\varphi(x)$ when there are some reasons to assume that $\varphi(x)$ can be approximated sufficiently accurately by a polynomial of the third degree on a given interval $[a, b]$. In this case the derivative of the approximating function is a polynomial of the second degree, and, consequently, the problem reduces to finding the roots of a quadratic trinomial in the interval $[a, b]$.

8.5.2. Let an interval $[a, b]$ containing the point of minimum of $\varphi(x)$ be known and let $\varphi(x) \simeq P_3(x)$ on $[a, b]$. Assume that $P_3'(x) = P_2(x) \simeq \varphi'(x)$ on the interval under consideration. It can easily be verified that

$P_2(x)$ vanishes at the point

$$\tilde{x} = b - \frac{(z + \varphi'(b) - w)\lambda}{\varphi'(b) + 2z + \varphi'(a)}, \tag{8.19}$$

where

$$\lambda = b - a, \quad z = \frac{3[\varphi(a) - \varphi(b)]}{\lambda} + \varphi'(a) + \varphi'(b),$$
$$w = \sqrt{z^2 - \varphi'(a)\varphi'(b)}.$$

8.5.3. The scheme of the method. It is assumed that the values $\varphi(a)$, $\varphi(b)$, $\varphi'(a)$, and $\varphi'(b)$ are known and that $\varphi'(a) < 0$, $\varphi'(b) > 0$. Also, let some numbers $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$ be given.

Step 1. If $\lambda \leqslant \varepsilon_1$, put $x^* \simeq b$ and stop the search. If otherwise, calculate $\tilde{x}$ by formula (8.19).

Step 2. Calculate $\varphi(\tilde{x})$ and $\varphi'(\tilde{x})$. If $|\varphi'(\tilde{x})| \leqslant \varepsilon_2$, put $x^* \sim \tilde{x}$ and stop the search. If otherwise, pass to step 3.

Step 3. If $\varphi'(\tilde{x}) < -\varepsilon_2$, put $a = \tilde{x}$, $\varphi(a) = \varphi(\tilde{x})$, $\varphi'(a) = \varphi'(\tilde{x})$ and pass to step 1. If $\varphi'(\tilde{x}) > \varepsilon_2$, put $b = \tilde{x}$, $\varphi(b) = \varphi(\tilde{x})$, $\varphi'(b) = \varphi'(\tilde{x})$ and pass to step 1.

8.6. Method of Tangents

8.6.1. The idea of the method. This method may prove very effective in the minimization of convex functions.

We shall assume that the function $\varphi(x)$ is convex and continuously differentiable on an interval $[a, b]$. Denote by $\varphi_l(x, y)$ the following function linear in x and defined at the point y:

$$\varphi_l(x, y) = \varphi(y) + \varphi'(y)(x - y).$$

The method consists in the following. One takes an arbitrary point $x_0 \in [a, b]$ and constructs the function $\varphi_l(x, x_0)$. To calculate the point x_1 the following problem is solved:

$$\min \xi$$

under the conditions

$$\xi \geqslant \varphi_l(x, x_0)$$
$$a \leqslant x \leqslant b.$$

Denote by ξ_1, x_1 the solution to this problem. To find the point x_2 the function $\varphi_l(x, x_1)$ is constructed and ξ_2, x_2, i.e. the solution to the problem

$$\min \xi$$

under the conditions

$$\xi \geqslant \varphi_l(x, x_0),$$
$$\xi \geqslant \varphi_l(x, x_1),$$
$$a \leqslant x \leqslant b$$

is calculated. Let the points $x_0, x_1, \ldots, x_{k-1}$ be calculated. To find x_k one solves the problem

$$\min \xi \tag{8.20}$$

under the conditions

$$\begin{aligned} &\xi \geqslant \varphi_l(x, x_0), \\ &\xi \geqslant \varphi_l(x, x_1), \\ &\cdots\cdots\cdots \\ &\xi \geqslant \varphi_l(x, x_{k-1}), \\ &a \leqslant x \leqslant b. \end{aligned} \tag{8.21}$$

Let

$$\psi_{k-1}(x) = \max\{\varphi_l(x, x_0), \varphi_l(x, x_1), \ldots, \varphi_l(x, x_{k-1})\} \tag{8.22}$$

or, which is the same,

$$\psi_{k-1}(x) = \max\{\psi_{k-2}(x), \varphi_l(x, x_{k-1})\}. \tag{8.23}$$

It can easily be shown that for $\varphi'(x_{k-1}) \neq 0$ the problem (8.20), (8.21) is equivalent to the problem of finding a point $x_k \in [a, b]$ such that

$$\psi_{k-1}(x_k) = \min_{x \in [a, b]} \psi_{k-1}(x). \tag{8.24}$$

It is seen from Fig. 8.1 that to calculate each point x_k it suffices to determine at each step two new vertices of the broken line $\psi_{k-1}(x)$ (in the figure the broken line ACB represents the graph of the function $\psi_1(x)$ and $ADEB$ represents the graph of the function $\psi_2(x)$).

8.6.2. Properties of the functions $\psi_{k-1}(x)$. The convexity of the function $\varphi(x)$ implies that

$$\varphi(x) \geqslant \varphi(y) + \varphi'(y)(x - y),$$

which yields the inequality

$$\varphi_l(x, x_l) \leqslant \varphi(x)$$

holding for all $i = 0, 1, \ldots, k-1$ and any $x \in [a, b]$.

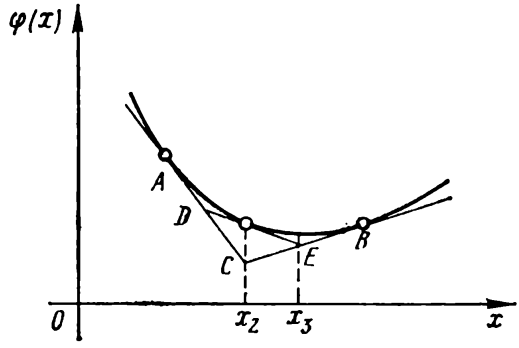


Fig. 8.1

Therefore,

$$\varphi(x) \geqslant \max\{\varphi_l(x, x_0), \varphi_l(x, x_1), \ldots, \varphi_l(x, x_{k-1})\}$$

and, consequently, by virtue of (8.22), the inequality

$$\psi_{k-1}(x) \leqslant \varphi(x) \tag{8.25}$$

is fulfilled for all $x \in [a, b]$. This and also the definition of the function $\varphi_l(x, y)$ and (8.22) imply that

$$\varphi(x_i) = \varphi_l(x_i, x_i) \leqslant \psi_{k-1}(x_i) \leqslant \varphi(x_i),$$

i.e.

$$\psi_{k-1}(x_i) = \varphi(x_i), \quad i = \overline{0, k-1}. \tag{8.26}$$

8.6.3. Justification of the convergence of the method. We shall assume that $\varphi'(x_k) \neq 0$ for all $k = 0, 1, \ldots$ since, if otherwise, x_k is a point of minimum, and the process is stopped at this step.

Theorem. *The method of tangents is convergent for a convex and continuously differentiable function $\varphi(x)$ on an interval $[a, b]$:*

$$\lim_{k \to \infty} \psi_k(x_{k+1}) = \varphi(x^*) = \min_{x \in [a, b]} \varphi(x).$$

If the point of minimum x^* is unique, then

$$\lim_{k\to\infty} x_k = x^*.$$

Proof. From (8.23)-(8.25) we find

$$\psi_{k-1}(x_k) = \min_{x\in[a,\,b]} \psi_{k-1}(x) \leqslant \psi_{k-1}(x_{k+1}) \leqslant \psi_k(x_{k+1}) = \min_{x\in[a,\,b]} \psi_k(x) \leqslant \psi_k(x^*) \leqslant \varphi(x^*).$$

Hence, the sequence $\{\psi_k(x_{k+1})\}$ is monotonically increasing and bounded from above, and therefore there exists

$$\psi^* = \lim_{k\to\infty} \psi_k(x_{k+1}).$$

The foregoing inequality implies that

$$\psi^* \leqslant \varphi(x^*). \tag{8.27}$$

We shall prove that $\psi^* = \varphi(x^*)$. Let

$$L = \max_{x\in[a,\,b]} |\varphi'(x)|.$$

We note that, in view of the convexity of the function $\varphi(x)$,

$$L = \max\{|\varphi'(a)|, |\varphi'(b)|\}$$

and the broken line $\psi_k(x)$ satisfies the Lipschitz condition with the constant L on the interval $[a, b]$.

Since $x_k \in [a, b]$ $(k = 0, 1, \ldots)$, there exists a convergent subsequence $\{x_{k_i}\}$:

$$\bar{x} = \lim_{i\to\infty} x_{k_i}.$$

From

$$\psi_{k_i-1}(x_{k_i}) = \min_{x\in[a,\,b]} \psi_{k_i-1}(x) \leqslant \psi_{k_i-1}(x_{k_i-1}),$$

by virtue of the Lipschitz condition for the function $\psi_{k_i-1}(x)$, we obtain

$$0 \leqslant \psi_{k_i-1}(x_{k_i-1}) - \psi_{k_i-1}(x_{k_i}) \leqslant L\,|x_{k_i-1} - x_{k_i}|.$$

One can assume that $k_1 < k_2 < \ldots < k_i \ldots$, then $k_{i-1} \leqslant k_i - 1$, and, consequently (see (8.26)),

$$\psi_{k_i-1}(x_{k_{i-1}}) = \varphi(x_{k_{i-1}}).$$

Therefore

$$0 \leqslant \varphi(x_{k_{i-1}}) - \psi_{k_i - 1}(x_{k_i}) \leqslant L \mid x_{k_{i-1}} - x_{k_i} \mid.$$

Passing to the limit for $i \to \infty$ we arrive at

$$0 \leqslant \varphi(\bar{x}) - \psi^* \leqslant 0,$$

whence it follows that $\varphi(\bar{x}) = \psi^*$. However, $\varphi(x^*) \leqslant \varphi(\bar{x})$, and therefore $\varphi(x^*) \leqslant \psi^*$. Thus, by virtue of (8.27),

$$\psi^* = \varphi(x^*).$$

The first part of the theorem is proved. The second assertion of the theorem is an obvious consequence of the property that any limiting point $\bar{x}$ of the sequence $\{x_k\}$ is such that

$$\varphi(\bar{x}) = \varphi(x^*) = \min_{x \in [a, b]} \varphi(x). \quad \triangle$$

8.6.4. The non-differentiable case. The method of tangents can also be used for one-dimensional minimization of convex functions not satisfying the condition of continuous differentiability. Since a convex function $\varphi(x)$ possesses left-hand and right-hand derivatives $\varphi'(y - 0)$ and $\varphi'(y + 0)$ at each interior point y of the interval $[a, b]$, the method of tangents can be modified for this case as well in the following way. As a linear function $\varphi_l(x, x_i)$ we take the following expression:

$$\varphi_l(x, x_i) = \varphi(x_i) + \gamma_i (x - x_i),$$

where γ_i is an arbitrary number belonging to the interval $[\varphi'(x_i - 0), \varphi'(x_i + 0)]$. In particular, γ_i can be found from the condition

$$| \gamma_i | = \min \{ | \varphi'(x_i - 0) |, | \varphi'(x_i + 0) | \}.$$

The further construction of the method remains unchanged. The convergence theorem remains true in this case as well.

8.6.5. The scheme of the method. The algorithm below makes it possible to determine the point of minimum of a convex and differentiable function $\varphi(x)$ on an interval $[a, b]$ with a preassigned accuracy ε.

Step 1 (nonstandard). Put $a_1 = a$, $b_1 = b$, calculate $\varphi(a_1)$, $\varphi(b_1)$, $\varphi'(a_1)$, $\varphi'(b_1)$, and pass to step 2 for $k = 1$.

Step 2. If $(b_k - a_k)/2 \leqslant \varepsilon$, stop the search and put $x^* \simeq \arg\min \{\varphi(a_k), \varphi(b_k)\}$.

If $(b_k - a_k)/2 > \varepsilon$, calculate*

$$x_k = \frac{[\varphi'(b_k)\, b_k - \varphi'(a_k)\, a_k] - [\varphi(b_k) - \varphi(a_k)]}{\varphi'(b_k) - \varphi'(a_k)},$$

$\varphi(x_k)$, $\varphi'(x_k)$. If $\varphi'(x_k) = 0$, then $x^* = x_k$, and the search is ended. If $\varphi'(x_k) < 0$, put $a_{k+1} = x_k$, $b_{k+1} = b_k$. If $\varphi'(x_k) > 0$, put $a_{k+1} = a_k$, $b_{k+1} = x_k$.

---

* It is readily seen that x_k is the point of intersection of the straight lines $y = \varphi(a_k) + \varphi'(a_k)(x - a_k)$ and $y = \varphi(b_k) + \varphi'(b_k)(x - b_k)$.

Chapter 9

RELAXATION METHODS OF SOLVING EXTREMAL PROBLEMS. UNCONDITIONAL MINIMIZATION METHODS

9.1. Convergence and Stability Problems. Relaxation Processes

9.1.1. Before beginning the solution of a specific extremal problem one should decide which of the minimization methods is preferable. To this end one must be able to compare the methods. One way is based on practice and a sufficient amount of experimental data, and no specialist using modern computers for solving applied problems can avoid it. Another way, which by no means excludes the former, is to compare the effectiveness of the methods applied to certain classes of problems. Here a priori characteristics of the methods are important including the amount of calculations, the rate of convergence, the stability of the methods relative to calculation errors, the computing time, etc. We have not yet posed the problem of exhaustive description of a method, however, it is clear that the question of convergence is one of its important components.

9.1.2. On the rate of convergence. Throughout this book the term the "rate of convergence" is understood in the classical sense, i.e. as a comparison of the deviation of the mth approximation from the exact solution with a certain function of the index m. Assume that for a definite class of problems method A provides a higher rate of convergence of successive approximations than method B. Then to make sure that this is the case for a specific extremal problem it is necessary to show that the problem belongs to that class. However, even this may not be a simple task. The matter is that the purpose of solving extremal problems is to understand the essence of the phenomenon in question and, in a sense, to make the best decision in controlling it. But, as a rule, the information about the phenomenon under study is incomplete, and, consequently, the initial data—the parameters of

the corresponding mathematical model—are also incomplete. Nevertheless, firstly, certain information about the problem is contained in its very formulation due to our knowledge of the "physical" aspect of the phenomenon and due to the available estimates for the accuracy of the input data. Secondly, while solving an extremal problem we obtain more information about it; hence, our knowledge that it belongs to a definite class is becoming more certain. Thus, the accumulated information and a priori characteristics make it possible to timely assess the advantages and shortcomings of a given method in application to a specific problem and, consequently, to introduce control elements into the search for the solution.

We note that, as a rule, estimates for the rate of convergence can be found only for some classes of functions under minimization. However, these estimates contain important information about the methods and provide ways for checking them experimentally for other classes of functions.

9.1.3. Assumptions. Below we consider various methods of solving extremal problems and study their convergence and stability for sufficiently smooth functions belonging to a class $C^{1,1}(X)$, namely, functions whose gradients satisfy the Lipschitz condition. Estimates for the rate of convergence of the methods under study are established under an additional assumption that the minimized function and the feasible region are convex.

9.1.4. On the stability of minimization methods. If for a given method there are a priori estimates for the order of the rate of convergence, for instance, allowing one to compare the deviation of the mth approximation from the exact solution with $O(1/m^{\alpha})$ (here the exponent α specifies the order of the rate of convergence), the method is said to be *stable relative to the choice of the direction of descent and the steplength* if it admits of an approximate determination of these parameters without changing the order of the rate of convergence. When a method is found to be convergent but there are no estimates for the rate of convergence, the method will be considered stable relative to the choice of the direction of descent and the steplength if it allows an approximate determination of

these parameters without violating the convergence of the minimization process. From what follows it will be clear that the methods of the steepest descent, gradient projection, conditional gradient, and feasible direction are stable.

9.1.5. The notion of a relaxation process. Consider the problem of minimizing a function $\varphi(\mathbf{x})$ on set a X. The procedure of constructing a sequence of points $\{\mathbf{x}_k\}$ will be called a *relaxation process* if

$$\mathbf{x}_k \in X \quad \text{and} \quad \varphi(\mathbf{x}_{k+1}) \leqslant \varphi(\mathbf{x}_k), \quad k = 0, 1, \ldots .$$

In what follows we shall always assume that the solution set of the original problem is nonempty. When investigating convergence of relaxation processes we shall also assume that $\{\mathbf{x}_k\}$ is an infinite sequence, and, consequently, its all points lie outside the solution set.

9.2. Auxiliary Material

9.2.1. We shall consider several auxiliary assertions as a necessary apparatus for investigating the convergence of relaxational minimization processes. In the present course Lemma 9.2.3 is basic for studying the convergence of minimization processes for functions of class $C^{1,1}(X)$. The material presented in Secs. 9.2.4-9.2.10 will be required for establishing the convergence of relaxation sequences to the set of points where necessary conditions for minimum are fulfilled, and Lemmas 9.2.14 and 9.2.15 are necessary for the derivation of estimates for the rate of convergence.

The assertions mentioned above equally apply to the study of both unconditional minimization methods and methods of minimizing functions under constraints.

9.2.2. The definition of the class $C^{1,1}(X)$. A differentiable function $\varphi(\mathbf{x})$ on a set X is said to be of class $C^{1,1}(X)$ if there is a constant $L > 0$ such that for any $\mathbf{x}, \mathbf{y} \in X$, $[\mathbf{x}, \mathbf{y}] \subset X$, the inequality

$$\| \varphi'(\mathbf{x}) - \varphi'(\mathbf{y}) \| \leqslant L \| \mathbf{x} - \mathbf{y} \|$$

is fulfilled.

9.2.3. Lemma. *If* $\varphi(\mathbf{x}) \in C^{1,1}(X)$, *then for any* $\mathbf{x}, \mathbf{y} \in X$ *such that* $[\mathbf{x}, \mathbf{y}] \subset X$ *the inequality*

$$\varphi(\mathbf{x}) - \varphi(\mathbf{y}) \geqslant \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \frac{L}{2} \| \mathbf{x} - \mathbf{y} \|^2$$

holds.

Proof. Using the condition $\varphi(\mathbf{x}) \in C^{1,1}(X)$ and the Cauchy-Bunyakovskiĭ inequality we obtain

$$\varphi(\mathbf{x}) - \varphi(\mathbf{y}) = \int_0^1 \langle \varphi'(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle \, d\tau$$

$$= \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle + \int_0^1 \langle \varphi'(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) - \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle \, d\tau$$

$$\geqslant \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \int_0^1 \| \varphi'(\mathbf{y} + \tau(\mathbf{x} - \mathbf{y})) - \varphi'(\mathbf{x}) \| \, \| \mathbf{x} - \mathbf{y} \| \, d\tau$$

$$\geqslant \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - L \| \mathbf{x} - \mathbf{y} \|^2 \int_0^1 (1 - \tau) \, d\tau$$

$$= \langle \varphi'(\mathbf{x}), \mathbf{x} - \mathbf{y} \rangle - \frac{L}{2} \| \mathbf{x} - \mathbf{y} \|^2. \quad \triangle$$

9.2.4. Consider some sets X and Y in an n-dimensional Euclidean space* such that $Y \subset X$ and $X \setminus Y \neq \varnothing$.

Let a function $f(\mathbf{x})$ on X be such that

$$f(\mathbf{y}) = 0 \text{ for all } \mathbf{y} \in Y. \tag{9.1}$$

Consider an arbitrary sequence $\{\mathbf{x}_k\} \subset X$ such that

$$f(\mathbf{x}_k) \to 0, \quad k \to \infty. \tag{9.2}$$

Given a number $\varepsilon \to 0$, we introduce the set

$$U_\varepsilon = \{\mathbf{x} \in X: \rho(\mathbf{x}, Y) \leqslant \varepsilon\},$$

where $\rho(\mathbf{x}, Y)$ is the distance from the point $\mathbf{x}$ to the set Y:

$$\rho(\mathbf{x}, Y) = \inf_{\mathbf{y} \in Y} \rho(\mathbf{x}, \mathbf{y}), \quad \rho(\mathbf{x}, \mathbf{y}) = \| \mathbf{x} - \mathbf{y} \|.$$

---

* All that is said below applies to an arbitrary metric space as well.

9.2.5. Theorem. *If the conditions* (9.1), (9.2) *are fulfilled and, moreover, for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that*

$$f(\mathbf{x}_k) \geqslant \delta \quad \text{for all} \quad \mathbf{x}_k \in X \setminus U_\varepsilon, \tag{9.3}$$

then

$$\rho(\mathbf{x}_k, Y) \to 0, \quad k \to \infty. \tag{9.4}$$

Proof. Assume that there are $\varepsilon > 0$ and a subsequence $\{\mathbf{x}_{k_i}\} \subset \{\mathbf{x}_k\}$ whose all elements lie outside the set U_ε: $\{x_{k_i}\} \subset X \setminus U_\varepsilon$. Then, according to (9.3), the inequality $f(\mathbf{x}_{k_i}) \geqslant \delta$ $(i = 0, 1, \ldots.)$ holds, which contradicts (9.2). $\triangle$

Discussion. We consider the application of Theorem 9.2.5 to the study of convergence of relaxation processes for minimizing a function $\varphi(\mathbf{x})$ on a set X.

Consider the set $X_0 = \{\mathbf{x} \in X: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$ and the set $X^* \subset X_0$ of points where the necessary conditions for local minimum of the function $\varphi(\mathbf{x})$ are fulfilled. Assume that we know a function $f(\mathbf{x})$ such that

$$f(\mathbf{x}) > 0 \quad \text{for all} \quad \mathbf{x} \in X_0 \setminus X^*,$$
$$f(\mathbf{x}^*) = 0 \quad \text{for all} \quad \mathbf{x}^* \in X^*,$$

and let the inequality

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant f(\mathbf{x}_k)$$

hold for all the points $\mathbf{x}_k \in X_0 \setminus X^*$. Then the monotonicity and the boundedness of the sequence $\{\varphi(\mathbf{x}_k)\}$ imply that $\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \to 0$ for $k \to \infty$; hence, $f(\mathbf{x}_k) \to 0$, $k \to \infty$. If $f(\mathbf{x})$ satisfies (9.3), i.e. for any $\varepsilon > 0$ there is $\delta = \delta(\varepsilon) > 0$ such that $f(\mathbf{x}) \geqslant \delta$ for all $\mathbf{x} \in X_0 \setminus U_\varepsilon$ (here $U_\varepsilon = \{\mathbf{x} \in X_0 \setminus X^*: \rho(\mathbf{x}, X^*) \leqslant \varepsilon\}$), then, based on Theorem 9.2.5, we arrive at $\lim\limits_{k \to \infty} \rho(\mathbf{x}_k, X^*) = 0$.

In methods of unconditional minimization of smooth functions we usually have the inequality $\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant C \| \varphi'(\mathbf{x}_k) \|^2$. Taking $f(\mathbf{x}) = \| \varphi'(\mathbf{x}) \|^2$ and applying Theorem 9.2.5 we prove the convergence of successive approximations to the set X^*.

9.2.6. Theorem. *If*

(1) $\varphi(\mathbf{x})$ *is a convex function of class* $C^{1,1}(E_n)$,

(2) $X^* = \{\mathbf{x}^*: \varphi'(\mathbf{x}^*) = 0\} \neq \varnothing$,

(3) *for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that* $\| \varphi'(\mathbf{x}) \| \geqslant \delta$ *for all* $\mathbf{x}$ *such that*

$$\rho(\mathbf{x}, X^*) = \| \mathbf{x} - \mathbf{p}^* \| \geqslant \varepsilon,$$

then any sequence

$$\{\mathbf{x}_k\} \subset X_0 = \{\mathbf{x}: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$$

satisfies

$$\rho(\mathbf{x}_k, X^*) = \| \mathbf{x}_k - \mathbf{p}_k^* \| \leqslant \eta < \infty, \quad k = 0, 1, \ldots.$$

Proof. Denote $\varphi^* = \min \varphi(\mathbf{x})$, $C = \varphi(\mathbf{x}_0) - \varphi^*$. Then $\varphi(\mathbf{x}_k) - \varphi^* \leqslant C < \infty$ $(k = 0, 1, \ldots)$. For an arbitrary $\varepsilon > 0$ and all indices k such that $\rho(\mathbf{x}_k, X^*) = \| \mathbf{x}_k - \mathbf{p}_k^* \| \geqslant \varepsilon$ we consider

$$\mathbf{y}_k = \frac{\varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|} \mathbf{x}_k + \left(1 - \frac{\varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|}\right) \mathbf{p}_k^*.$$

Since $\mathbf{p}_k^*$ is the projection of the point $\mathbf{x}_k$ on the convex and closed set X^* of points of minimum of the function $\varphi(\mathbf{x})$ and $\mathbf{y}_k \in [\mathbf{p}_k^*, \mathbf{x}_k]$, we have $\rho(\mathbf{y}_k, X^*) = \| \mathbf{y}_k - \mathbf{p}_k^* \| = \varepsilon$, and, by virtue of the assumption (3), $\| \varphi'(\mathbf{y}_k) \| \geqslant \delta$.

Using Lemma 9.2.3 we find

$$\text{(A)} \quad \varphi(\mathbf{y}_k) - \varphi^* \geqslant \varphi(\mathbf{y}_k) - \varphi\left(\mathbf{y}_k - \frac{\delta}{L} \frac{\varphi'(\mathbf{y}_k)}{\| \varphi'(\mathbf{y}_k) \|}\right)$$
$$\geqslant \frac{\delta}{L} \left\langle \varphi'(\mathbf{y}_k), \frac{\varphi'(\mathbf{y}_k)}{\| \varphi'(\mathbf{y}_k) \|} \right\rangle - \frac{1}{2} L \frac{\delta^2}{L^2} \geqslant \frac{1}{2} \frac{\delta^2}{L} > 0.$$

Assume that the theorem is false, i.e. there is a sequence $\{\mathbf{x}_k\}$ for which

(B) $\| \mathbf{x}_k - \mathbf{p}_k^* \| \to \infty$, $k \to \infty$.

The convexity of the function $\varphi(\mathbf{x})$ and the relation $\varphi^* = \varphi(\mathbf{p}_k^*)$ imply

$$\varphi(\mathbf{y}_k) \leqslant \frac{\varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|} \varphi(\mathbf{x}_k) + \left(1 - \frac{\varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|}\right) \varphi^*$$
$$= \varphi^* + \frac{\varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|} (\varphi(\mathbf{x}_k) - \varphi^*) \leqslant \varphi^* + \frac{C \cdot \varepsilon}{\| \mathbf{x}_k - \mathbf{p}_k^* \|}.$$

Taking the assumption (B) into consideration we arrive at $\lim_{k \to \infty} \varphi(\mathbf{y}_k) = \varphi^*$, which contradicts the inequality (A). Δ

The above theorem plays an important role in justifying estimates for the rate of convergence of minimization methods for convex functions since these estimates are essentially based on the condition that $\| \mathbf{x}_k - \mathbf{p}_k^* \|$ is bounded.

9.2.7. We now investigate the properties of the sets X and Y and the function $f(\mathbf{x})$ that guarantee the fulfilment of the condition (9.3).

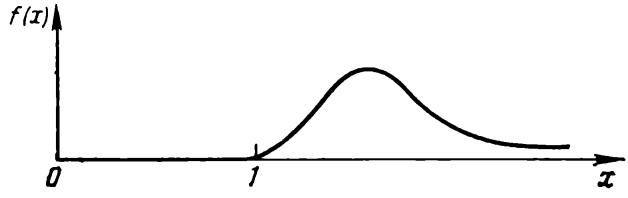


Fig. 9.1

Theorem. *If the set X is bounded and closed and the function $f(\mathbf{x})$ on X is continuous and such that*

$$f(\mathbf{y}) = 0 \text{ for all } \mathbf{y} \in Y,$$
$$f(\mathbf{x}) > 0 \text{ for all } \mathbf{x} \in X \setminus Y \neq \varnothing,$$

then the condition (9.3) *holds.*

Proof. Assume that there are $\varepsilon > 0$ and a sequence $\{\mathbf{x}_k\} \in X \setminus U_\varepsilon$ such that $f(\mathbf{x}_k) \to 0$, $k \to \infty$. In view of the boundedness of the set X, the sequence $\{\mathbf{x}_k\}$ contains a convergent subsequence $\{\mathbf{x}_{k_i}\}$, and we have $\lim\limits_{i \to \infty} \mathbf{x}_{k_i} = \mathbf{x} \in X \setminus Y$. However, $f(\mathbf{x}) = 0$, and, consequently, $\mathbf{x} \in Y$, which is a contradiction. $\triangle$

Remark. In case the set X is unbounded, the condition (9.3) may be violated. Figure 9.1 represents the graph of a one-dimensional function on the semiaxis $x \geqslant 0$ (the set X) vanishing on the closed interval $[0, 1]$ (the set Y) for which (9.3) does not hold.

9.2.8. Finally, we indicate one more property which will further be used in both investigating the existence of a global minimum and checking the condition (9.3).

Let $\mathbf{z}$ be an arbitrary point belonging to a set X.

If the set X is unbounded and a function $f(\mathbf{x})$ on X possesses the property $f(\mathbf{x}_k) \to +\infty$, $k \to \infty$ for any se-

quence $\{\mathbf{x}_k\} \subset X$ *such that* $\| \mathbf{x}_k \| \to \infty$, $k \to \infty$, *then the set* $X(\mathbf{z}) = \{\mathbf{x} \in X: f(\mathbf{x}) \leqslant f(\mathbf{z})\}$ *is bounded.*

9.2.9. In conclusion we find conditions under which the convergence of a sequence $\{\mathbf{x}_k\}$ to a set implies its convergence to a point belonging to the set.

Theorem. *If the set* Y $(Y \subset X)$ *of limiting points of a sequence* $\{\mathbf{x}_k\} \subset X$ *is finite* $(Y = \{\mathbf{y}_1, \ldots, \mathbf{y}_m\})$, $\lim\limits_{k\to\infty} \| \mathbf{x}_k - \mathbf{x}_{k+1} \| = 0$, *and* $\lim\limits_{k\to\infty} \rho(\mathbf{x}_k, Y) = 0$, *then the sequence is convergent.*

Proof. Denote $U_\delta(\mathbf{y}) = \{\mathbf{x} \in X: \rho(\mathbf{x}, \mathbf{y}) \leqslant \delta\}$. If $m = 1$, the assertion is trivial.

Let $m > 1$. We denote $\delta = \min\limits_{i\neq j} \| \mathbf{y}_i - \mathbf{y}_j \| > 0$ $(i, j = 1, \ldots, m)$. It follows from the conditions of the theorem that there is $k_0 \geqslant 0$ such that $\mathbf{x}_k \in \bigcup\limits_{i=1}^{m} U_{\delta/4}(\mathbf{y}_i)$ and $\| \mathbf{x}_k - \mathbf{x}_{k+1} \| \leqslant \delta/4$ for all $k \geqslant k_0$. If $\mathbf{x}_{k_1} \in U_{\delta/4}(\mathbf{y}_1)$ for some $k_1 \geqslant k_0$, then for any $i \geqslant 2$ we have $\| \mathbf{y}_i - \mathbf{x}_{k_i+1} \| = \| (\mathbf{y}_i - \mathbf{y}_1) + (\mathbf{y}_1 - \mathbf{x}_{k_i}) + (\mathbf{x}_{k_i} - \mathbf{x}_{k_i+1}) \| \geqslant \| \mathbf{y}_i - \mathbf{y}_1 \| - (\| \mathbf{y}_1 - \mathbf{x}_{k_i} \| + \| \mathbf{x}_{k_i} - \mathbf{x}_{k_i+1} \|) \geqslant \delta - 2(\delta/4) = \delta/2$, i.e. $\mathbf{x}_{k_i+1}$ does not belong to any of the neighborhoods of the points $\mathbf{y}_2, \ldots, \mathbf{y}_m$, and, consequently, $\mathbf{x}_{k_i+1} \in U_{\delta/4}(\mathbf{y}_1)$. By induction, it follows that $\mathbf{x}_k \in U_{\delta/4}(\mathbf{y}_1)$ for all $k \geqslant k_1$. Since $\rho(\mathbf{x}_k, Y) \to 0$, we have $\lim\limits_{k\to\infty} \mathbf{x}_k = \mathbf{y}_1$. Thus, the entire sequence $\{\mathbf{x}_k\}$ converges to one of the points belonging to the finite set Y. $\triangle$

9.2.10. Theorem. *If* $f(\mathbf{x})$ *is a convex and differentiable function on a convex set* X *and the set* $Y = \operatorname{Arg\,min}\{f(\mathbf{x}): \mathbf{x} \in X\}$ *is bounded, then*

$$\forall \varepsilon > 0\ \exists \delta = \delta(\varepsilon) > 0:\ \forall \mathbf{x} \in X \setminus U_\varepsilon(Y): \| f'(\mathbf{x}) \| \geqslant \delta,$$

where $U_\varepsilon(Y) = \{\mathbf{x}: \rho(\mathbf{x}, Y) \leqslant \varepsilon\}$.

Proof. We denote by G the boundary of the set $U_\varepsilon(Y)$. The set $G_1 = G \cap X$ is bounded and closed. It is obvious that for any fixed $y \in Y$ we have $\zeta = \min\limits_{\mathbf{z}\in G_1} (f(\mathbf{z}) - f(\mathbf{y})) > 0$. Let $\mathbf{x}$ be an arbitrary point belonging to the set $X \setminus U_\varepsilon(Y)$ and let $\mathbf{y} \in Y$. It is clear that in the line segment $[\mathbf{y}, \mathbf{x}]$ there is a point $\mathbf{z} \in G$, and we have $\| \mathbf{z} - \mathbf{y} \| = \eta < \infty$ since the set $X \cap U_\varepsilon(Y)$ is bounded.

The convex function of one variable $F(\lambda) = f(\mathbf{z} + \lambda(\mathbf{z} - \mathbf{y}))$ satisfies the inequality $F'(0) \leqslant F'(\lambda)$ for all $\lambda \geqslant 0$, and, consequently, $\langle f'(\mathbf{z}), \mathbf{z} - \mathbf{y}\rangle \leqslant \langle f'(\mathbf{x}), \mathbf{z} - \mathbf{y}\rangle$. Since for any $\mathbf{x} \in X \setminus U_\varepsilon(Y)$ there holds

$$\zeta \leqslant f(\mathbf{z}) - f(\mathbf{y}) \leqslant \langle f'(\mathbf{z}), \mathbf{z} - \mathbf{y}\rangle \leqslant \langle f'(\mathbf{x}), \mathbf{z} - \mathbf{y}\rangle \leqslant \| f'(\mathbf{x}) \| \, \| \mathbf{z} - \mathbf{y} \| \leqslant \eta \| f'(\mathbf{x}) \|,$$

for $\delta = \zeta/\eta$ we obtain the assertion of the theorem. $\triangle$

Discussion. Assume that for a relaxational minimization process (see 9.2.6) it is proved that $\lim\limits_{k\to\infty} \rho(\mathbf{x}_k, X^*) = 0$. If the set X^* of points of local minima of the function $\varphi(\mathbf{x})$ is finite and the condition $\lim\limits_{k\to\infty} \| \mathbf{x}_k - \mathbf{x}_{k+1} \| = 0$ is fulfilled, then Theorem 9.2.9 implies that the entire sequence converges to a point $\mathbf{x}^*$ of local minimum. Naturally, this limiting point of the minimizing sequence $\{\mathbf{x}_k\}$ depends on the choice of the initial point $\mathbf{x}_0$. We note that the relation $\| \mathbf{x}_k - \mathbf{x}_{k+1} \| \to 0$, $k \to \infty$, is fulfilled for many relaxational minimization processes. Below we consider some conditions guaranteeing the fulfilment of this relation.

9.2.11. Linearly non-constant functions. A function $\varphi(\mathbf{x})$ will be called *linearly non-constant* on a set X if there are no two different points $\mathbf{x}, \mathbf{y} \in X$ such that $[\mathbf{x}, \mathbf{y}] \subset X$ and $\varphi(\mathbf{z}) = \varphi(\mathbf{x})$ $\forall \mathbf{z} \in (\mathbf{x}, \mathbf{y}]$.

Examples of linearly non-constant functions are strictly convex functions and also functions satisfying the condition

$$\varphi(\mathbf{z}) < \max\{\varphi(\mathbf{x}), \varphi(\mathbf{y})\} \quad \forall \mathbf{z} \in (\mathbf{x}, \mathbf{y}],$$

the so-called *strictly quasi-convex functions.*

9.2.12. Strongly lowering sequences. A sequence $\{\mathbf{x}_k\} \subset X$ is said to be *strongly lowering* for a function $\varphi(\mathbf{x})$ on the set X if

$$[\mathbf{x}_k, \mathbf{x}_{k+1}] \subset X \text{ and } \varphi(\mathbf{x}_k) \geqslant \varphi(\mathbf{z}) \geqslant (\mathbf{x}_{k+1}) \quad \forall \mathbf{z} \in [\mathbf{x}_k, \mathbf{x}_{k+1}].$$

In the majority of minimization methods considered in the subsequent sections we deal with sequences possessing the indicated property.

9.2.13. Theorem. *If a function $\varphi(\mathbf{x})$ is continuous and linearly non-constant on a closed and bounded set X, then*

any strongly lowering sequence $\{\mathbf{x}_k\}$ *satisfies the condition*

$$\lim_{k\to\infty} \| \mathbf{x}_k - \mathbf{x}_{k+1} \| = 0.$$

Proof. Assume that there is a subsequence $\{\mathbf{x}_{k_i}\}$ such that

$$\| \mathbf{x}_{k_i+1} - \mathbf{x}_{k_i} \| \geqslant \varepsilon > 0 \quad \forall i \geqslant 0.$$

Since $\{\mathbf{x}_{k_i}\}$ is bounded, we can assume, without changing the notation, that

$$\lim_{i\to\infty} \mathbf{x}_{k_i} = \mathbf{x} \quad \text{and} \quad \lim_{i\to\infty} \mathbf{x}_{k_i+1} = \mathbf{y}.$$

Then $\| \mathbf{x} - \mathbf{y} \| \geqslant \varepsilon > 0$, and the closedness of the set X and the condition $[\mathbf{x}_k, \mathbf{x}_{k+1}] \subset X$ imply $[\mathbf{x}, \mathbf{y}] \subset X$. Since the sequence $\{\varphi(\mathbf{x}_k)\}$ is monotonic and bounded, we obtain

$$\lim_{k\to\infty} (\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})) = 0,$$

and, consequently, $\varphi(\mathbf{x}) = \varphi(\mathbf{y})$, whence, by the condition $\varphi(\mathbf{x}_k) \geqslant \varphi(\mathbf{z}) \geqslant \varphi(\mathbf{x}_{k+1})$, it follows that

$$\varphi(\mathbf{x}) = \varphi(\mathbf{z}) = \varphi(\mathbf{y}) \quad \forall \mathbf{z} \in [\mathbf{x}, \mathbf{y}],$$

which contradicts the assumption that the function $\varphi(\mathbf{x})$ is linearly non-constant. △

Taking into account what was said in the foregoing section, we see that the theorem we have proved asserts that many minimization methods generate sequences of contracting intervals $[\mathbf{x}_k, \mathbf{x}_{k+1}]$.

9.2.14. Lemma. *If a number sequence* $\{\mu_k\}$ *is such that* $\mu_k - \mu_{k+1} \geqslant \tau_k \mu_k^2$, $\mu_k > 0$, $\tau_k \geqslant 0$, $k = 0, 1, \ldots$, then

$$\mu_m \leqslant \mu_0 \left[1 + \mu_0 \sum_{k=0}^{m-1} \tau_k\right]^{-1}, \quad m = 1, 2, \ldots$$

Proof. The conditions of the lemma imply that

$$\frac{\mu_k}{\mu_{k+1}} \geqslant 1 + \tau_k \frac{\mu_k^2}{\mu_{k+1}} \geqslant 1.$$

Therefore

$$\frac{1}{\mu_{k+1}} - \frac{1}{\mu_k} = \frac{\mu_k - \mu_{k+1}}{\mu_k \mu_{k+1}} \geqslant \tau_k \frac{\mu_k}{\mu_{k+1}} \geqslant \tau_k.$$

Summing this inequality over k we find

$$\sum_{k=0}^{m-1}\left[\frac{1}{\mu_{k+1}}-\frac{1}{\mu_k}\right]=\frac{1}{\mu_m}-\frac{1}{\mu_0}\geqslant\sum_{k=0}^{m-1}\tau_k,$$

whence the desired estimate follows. $\triangle$

9.2.15. Lemma. *If a number sequence* $\{\mu_k\}$ *is such that* $\mu_k-\mu_{k+1}\geqslant\tau_k\mu_k$, $\mu_k>0$, $\tau_k\geqslant 0$, $k=0, 1, \ldots$ *then*

$$\mu_m\leqslant\mu_0\exp\left\{-\sum_{k=0}^{m-1}\tau_k\right\},\quad m=1, 2, \ldots$$

The proof obviously follows from the condition $\mu_k>0$ since $0<1-\tau_k\leqslant 1$, and, consequently,

$$\mu_m\leqslant(1-\tau_{m-1})\,\mu_{m-1}\leqslant\mu_0\prod_{k=0}^{m-1}(1-\tau_k)$$

$$=\mu_0\exp\left\{\sum_{k=0}^{m-1}\ln(1-\tau_k)\right\}\leqslant\mu_0\exp\left\{-\sum_{k=0}^{m-1}\tau_k\right\}.\quad\triangle$$

9.2.16. Theorem. *Let a function* $f(\mathbf{x})$ *be convex and differentiable on a convex set* X *and let the set* $Y=\operatorname{Arg}\min\{f(\mathbf{x}):\mathbf{x}\in X\}$ *be bounded. If a sequence* $\{\mathbf{x}_k\}\subset X$ *is such that* $\|f'(\mathbf{x}_k)\|\to 0$ $(k\to\infty)$, *then* $\lim\limits_{k\to\infty}\rho(\mathbf{x}_k, Y)=0$.

Proof. We first prove that the sequence $\{\mathbf{x}_k\}$ is bounded. Assume that $\|\mathbf{x}_k\|\to\infty$. Then there is $\varepsilon>0$ such that the set $X\setminus U_\varepsilon(Y)$ is nonempty and an index k_0 such that $\mathbf{x}_k\in X\setminus U_\varepsilon(Y)$ for all $k\geqslant k_0$. By Theorem 9.2.10, in this case there is $\delta=\delta(\varepsilon)>0$ such that $\|f'(\mathbf{x}_k)\|\geqslant\delta$ for all $k\geqslant k_0$. We have thus arrived at a contradiction.

We denote by $\mathbf{p}_k$ the projection of the point $\mathbf{x}_k$ on the set Y. Since the set Y and the sequence $\{\mathbf{x}_k\}$ are bounded, there is a number $\eta>0$ such that $\|\mathbf{x}_k-\mathbf{p}_k\|\leqslant\eta<\infty$. The function $f(\mathbf{x})$ being convex, we have

$$0\leqslant f(\mathbf{x}_k)-f(\mathbf{y})=f(\mathbf{x}_k)-f(\mathbf{p}_k)\leqslant\langle f'(\mathbf{x}_k),\,\mathbf{x}_k-\mathbf{p}_k\rangle$$
$$\leqslant\|f'(\mathbf{x}_k)\|\,\|\mathbf{x}_k-\mathbf{p}_k\|\leqslant\eta\,\|f'(\mathbf{x}_k)\|.$$

From the relation $f(\mathbf{x}_k)\to f(\mathbf{y})$ $(k\to\infty)$ and the boundedness of the sequence $\{\mathbf{x}_k\}$ it follows that $\lim\limits_{k\to\infty}\rho(\mathbf{x}_k, Y)=0$. $\triangle$

9.3. Theorems on Estimates

9.3.1. In this section we consider estimates for the rate of convergence of relaxational minimization processes of a general type irrespective of their specific realization. It is important that the estimates are valid only for convex programming problems.

Thus, throughout the section it is assumed that the function $\varphi(x)$ under minimization and the feasible region X are convex.

We denote

$$x^* = \arg\min\{\varphi(x): x \in X\}, \quad X^* = \{x^* \in X\},$$
$$\varphi^* = \varphi(x^*), \quad \mu_k = \varphi(x_k) - \varphi^*, \quad \rho(x_k, X^*) = \| x_k - p_k^* \|,$$

where p_k^* is the projection of the point x_k on the set X^*.

The convexity of the function $\varphi(x)$ implies the inequality

$$0 < \mu_k \leqslant \langle \varphi'(x_k), x_k - p_k^* \rangle \tag{9.5}$$

holding for all the points x_k owing to the assumption that

$$\varphi(x_k) \neq \varphi^*.$$

The majority of studies of the rate of convergence of relaxational minimization methods are based on the following theorem on estimates.

9.3.2. Theorem. *If*

(1) *the function* $\varphi(x)$ *is convex and differentiable on a convex and closed set* X,

(2) *the set* X^* *is nonempty*,

(3) $\{x_k\}$ *is a relaxation sequence*,

then the inequality

$$\varphi(x_m) - \varphi^* \leqslant \mu_0 \left[1 + \mu_0 \sum_{k=0}^{m-1} \frac{\varphi(x_k) - \varphi(x_{k+1})}{\langle \varphi'(x_k), x_k - p_k^* \rangle^2} \right]^{-1} \tag{9.6}$$

holds.

Proof. Since (see (9.5))

$$\mu_k - \mu_{k+1} = \varphi(x_k) - \varphi(x_{k+1})$$
$$= \frac{\varphi(x_k) - \varphi(x_{k+1})}{\langle \varphi'(x_k), x_k - p_k^* \rangle^2} \langle \varphi'(x_k), x_k - p_k^* \rangle^2$$
$$\geqslant \frac{\varphi(x_k) - \varphi(x_{k+1})}{\langle \varphi'(x_k), x_k - p_k^* \rangle^2} \mu_k^2,$$

the application of Lemma 9.2.14 to this inequality results in (9.6). Δ

The estimate (9.6) is of a non-constructive character since its right-hand side involves the unknown projections $\mathbf{p}_k^*$. However, the derivation of estimates for the rate of convergence of specific methods for solving convex programming problems is essentially based on the above theorem.

It is obvious that if the series on the right-hand side of the inequality (9.6) diverges for $m \to \infty$, then the relaxation process is weakly convergent (i.e. the convergence $\varphi(\mathbf{x}_m) \to \varphi^*$ relative to the functional takes place). In this case the strong convexity of the function $\varphi(\mathbf{x})$ guarantees (see 2.5.5) the strong convergence of the process (i.e. the convergence of the sequence $\{\mathbf{x}_k\}$ to the solution $\mathbf{x}^*$). Hence, to investigate the convergence of relaxation processes using formula (9.6) one should be able to estimate from below the ratio

$$\frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k^* \rangle^2}.$$

We note that $\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k^* \rangle \to 0$ for $\varphi(\mathbf{x}_k) \to \varphi^*$.

For certain relaxation methods this fact makes it possible to find a constant $C > 0$ not depending on the number of iterations k such that

$$\frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k^* \rangle^2} \geqslant C$$

and thus to derive estimates of the form of $\varphi(\mathbf{x}_m) - \varphi^* \asymp O(1/m)$.

9.3.3. Theorem. *If*

(1) $X = E_n$,

(2) *the convex function* $\varphi(\mathbf{x})$ *is differentiable,*

(3) *the set* $X_0 = \{\mathbf{x}:\ \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$ *is bounded:* $\operatorname{diam} X_0 = \eta(\mathbf{x}_0) = \eta < \infty$,

(4) $\{\mathbf{x}_k\}$ *is a relaxation sequence, then the estimate*

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \mu_0 \left[1 + \mu_0 \frac{1}{\eta^2} \sum_{k=0}^{m-1} \frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\| \varphi'(\mathbf{x}_k) \|^2} \right]^{-1},$$

$$m = 1,\ 2,\ \ldots, \qquad (9.7)$$

takes place, where $\mu_0 = \varphi(\mathbf{x}_0) - \varphi^*$.

The theorem is an obvious consequence of (9.6) since

$$\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k^* \rangle^2 \leqslant \eta^2 \| \varphi'(\mathbf{x}_k) \|^2. \ \triangle$$

Remark. If there is an index m_0 such that $\varphi(\mathbf{x}_{m_0}) > \varphi(\mathbf{x}_{m_0+1})$, then (9.7) implies the inequality

$$\varphi(\mathbf{x}_m) - \varphi^* < \eta^2 \left[\sum_{k=0}^{m-1} \frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\| \varphi'(\mathbf{x}_k) \|^2} \right]^{-1},$$

$$m = m_0 + 1,\ m_0 + 2,\ \ldots \quad (9.8)$$

For convex functions of class $C^{1,1}(E_n)$ the condition that the set X_0 is bounded can be weakened by replacing it by the conditions of Theorem 9.2.6.

9.3.4. Theorem. *If*

(1) $X = E_n$,

(2) *the convex function* $\varphi(\mathbf{x})$ *is of class* $C^{1,1}(E_n)$,

(3) $X^* \neq \varnothing$,

(4) *for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that* $\| \varphi'(\mathbf{x}) \| \geqslant \delta$ *for all* $\mathbf{x}$ *satisfying*

$$\rho(\mathbf{x}, X^*) = \| \mathbf{x} - \mathbf{p}^* \| \geqslant \varepsilon,$$

(5) $\{\mathbf{x}_k\}$ *is a relaxation sequence,*

then the estimate (9.7) *holds.*

Proof. The conditions (1)-(4) are those enumerated in Theorem 9.2.6 which implies the inequality $\| \mathbf{x}_k - \mathbf{p}_k^* \| \leqslant \eta < \infty$ on which (9.7) is based. $\triangle$

9.3.5. Theorem. *If* $X = E_n$, *the function* $\varphi(\mathbf{x})$ *is strongly convex and differentiable, and* $\{\mathbf{x}_k\}$ *is a relaxation sequence, then the inequalities*

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \mu_0 \exp\left\{ -\rho \sum_{k=0}^{m-1} \frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\| \varphi'(\mathbf{x}_k) \|^2} \right\},$$

$$m = 1,\ 2,\ \ldots, \quad (9.9)$$

$$\| \mathbf{x}_m - \mathbf{x}^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp\left\{ -\rho \sum_{k=0}^{m-1} \frac{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})}{\| \varphi'(\mathbf{x}_k) \|^2} \right\},$$

$$m = 1,\ 2,\ \ldots, \quad (9.10)$$

hold, where ρ *a parameter of strong convexity.*

The proof obviously follows from the properties of strongly convex functions (see 2.5.5):

$$\mu_k \leqslant \frac{1}{\rho} \| \varphi'(x_k) \|^2,$$

$$\| x_k - x^* \|^2 \leqslant \frac{2}{\rho} [\varphi(x_k) - \varphi(x^*)]$$

and from Lemma 9.2.15:

$$\mu_k - \mu_{k+1} = \frac{\varphi(x_k) - \varphi(x_{k+1})}{\| \varphi'(x_k) \|^2} \| \varphi'(x_k) \|^2$$

$$\geqslant \rho \frac{\varphi(x_k) - \varphi(x_{k+1})}{\| \varphi'(x_k) \|^2} \mu_k.$$

9.3.6. The estimates (9.8)-(9.10) allow one to accumulate, during the computation process, information about the rate of convergence of $\varphi(x_m)$ to φ^* or, in the case of strong convexity, of x_m to x^*. We note that as long as the order of magnitude of the ratio

$$\frac{\varphi(x_k) - \varphi(x_{k+1})}{\| \varphi'(x_k) \|^2}$$

is not lower than $O(1/k)$, the relaxational minimization process remains convergent. In applications the following technique is used. One sets the parameter $\varepsilon > 0$, e.g. the one corresponding to the calculation accuracy chosen in the given stage, and the process is continued until the inequality

$$\frac{\varphi(x_k) - \varphi(x_{k+1})}{\| \varphi'(x_k) \|^2} \geqslant \varepsilon$$

is violated. However, the moment the situation where

$$\frac{\varphi(x_k) - \varphi(x_{k+1})}{\| \varphi'(x_k) \|^2} < \varepsilon$$

sets in and further application of the chosen algorithm leaves no hope that the situation may change, the iterative process is stopped. Minimization may go on either with a change in the algorithm parameters (e.g. with an increase of the calculation accuracy) or by choosing another method more effective in the case under consideration.

9.4. Descent Methods

9.4.1. The general scheme. The descent methods of solving unconditional minimization problems differ either in the choice of the direction of descent or in the way the motion along the latter is carried out. This makes it possible to write down the general scheme for the descent methods and investigate the related convergence and stability problems.

Thus, we consider the solution of the problem of minimizing a function $\varphi(\mathbf{x})$ on the entire space E_n. The descent methods consist in the following procedure of constructing the sequence $\{\mathbf{x}_k\}$. Generally, as an initial approximation an arbitrary point $\mathbf{x}_0 \in E_n$ is taken. The successive approximations $\mathbf{x}_1, \mathbf{x}_2, \ldots$ are constructed according to the following scheme:

(1) a direction of descent $-\mathbf{s}_k$ at the point $\mathbf{x}_k$ is selected,

(2) the $(k+1)$th approximation is found by the formula

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k. \tag{9.11}$$

9.4.2. Choice of the direction of descent. As was already said, the descent methods mainly differ in the choice of the direction of descent, and in the relaxation methods the direction $-\mathbf{s}_k$ is always chosen so that the inequality $\varphi(\mathbf{x}_{k+1}) \leqslant \varphi(\mathbf{x}_k)$ is fulfilled for at least small values of the parameter β_k. Since, as a rule, in actual problems we have information only about the behavior of the function $\varphi(\mathbf{x})$ in a local neighborhood of the point $\mathbf{x}_k$, the decrease of the function when passing from the point $\mathbf{x}_k$ in the direction $-\mathbf{s}_k$ is guaranteed by the relation

$$\langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle > 0.$$

In the majority of descent methods the choice of the direction is due to the computation of the gradient of the function under minimization at the point $\mathbf{x}_k$ (exceptions are, for instance, coordinate-wise descent methods and many of the random descent methods). The procedure of computing $\varphi'(\mathbf{x}_k)$ is extremely laborious. Practically, the gradient is usually calculated approximately using, for instance, various finite-difference schemes, which involves the calculation of the values of the function

$\varphi(\mathbf{x})$ at at least $n + 1$ points in each iteration. We note that the calculation of the values of the function $\varphi(\mathbf{x})$ is mostly connected with considerable difficulties, particularly when the values are found in an experiment.

The question of what method of selecting the direction of descent is preferable when solving a specific problem has no unambiguous answer unless we deal with a well-known test problem. The matter is that the "ideal" direction at the point $\mathbf{x}_0$ is the direction toward the point $\mathbf{x}^*$: $\mathbf{s} = \mathbf{x}_0 - \mathbf{x}^*$. In this sense all the other directions are worse. Assume that we have definite information about the properties of the minimized function; for instance, the function is convex and sufficiently smooth. However, even in this case it is rather difficult to answer the question since in the majority of descent methods a priori estimates for the rate of convergence have the form $\varphi(\mathbf{x}_m) - \varphi^* \asymp O(1/m)$, owing to which the estimates do not allow to give preference to a certain method. It is this stage where the accumulated experience becomes of major importance. For example, at the initial stage of minimization methods not requiring much effort when calculating the direction of descent should be chosen, e.g. coordinate-wise descent methods. When the process of minimizing $\varphi(\mathbf{x})$ is substantially decelerated, one passes to methods involving the computation of gradient, the so-called gradient descent methods. Further, if the successive approximations lead to a neighborhood of a point of minimum but the accuracy is still insufficient, some additional considerations are involved. For instance, if the behavior of the given convex function in the neighborhood of the point of minimum resembles that of a quadratic function, it is preferable to choose one of the conjugate directions methods since it is known that these methods minimize quadratic functions after a finite number of iterations.

What has been said should be regarded rather as a summary of the experience gained than as a "rule" for selecting the directions. The experience, intuition, and skill of a person dealing with numerical solution of real problems still play a very important role. Undoubtedly, the absence of a regular method for selecting directions hampers the procedure of solving extremal problems. The

only consolation is that this "blank spot" leaves a wide area for further studies.

9.4.3. Choosing the steplength. The parameter β_k determines, to within the factor $\| s_k \|$, the distance from the point x_k to the point x_{k+1}. Here we shall keep to the widespread terminology and call β_k a *steplength* or, simply a *step*.

When choosing the parameter β_k in relaxational minimization methods, the primary aim is to guarantee the fulfilment of the inequality $\varphi(x_{k+1}) < \varphi(x_k)$.

Naturally, the amount of calculations for finding β_k should be compatible with that for $-s_k$. If the choice of direction requires little effort, then, as a rule, the same must be true for finding β_k. One of the elementary methods implemented in a great number of standard minimization method programs is the duplication method.

9.4.4. The duplication method for choosing the step β_k. We take $\beta_k = \beta_{k-1}$. If this results in $\varphi(x_{k+1}) < \varphi(x_k)$, we either pass to the next $(k+2)$ th iteration or put $\beta_k = 2\beta_{k-1}$. If the calculated value of $\varphi(x)$ is less than the foregoing value, the duplication process can be continued until the decrease stops. If $\varphi(x_{k+1}) \geqslant \varphi(x_k)$, we take $\beta_k = \frac{1}{2}\beta_{k-1}$. If $\varphi(x_k - \frac{1}{2}\beta_{k-1}s_k) < \varphi(x_k)$, we put $x_{k+1} = x_k - \frac{1}{2}\beta_{k-1}s_k$ and pass to the next $(k+2)$th iteration. In case $\varphi(x_k - \frac{1}{2}\beta_{k-1}s_k) \geqslant \varphi(x_k)$, we take $\beta_k = \beta_{k-1}/4$, etc.

9.4.5. Choosing the step β_k by one-dimensional minimization methods. When the computation of s_k requires much effort, the parameter β_k is usually calculated by means of one-dimensional minimization methods in application to the function $\psi(\beta) = \varphi(x_k - \beta s_k)$ (see Chap. 8). In this case the accuracy in the calculation of a point of minimum of the function $\psi(\beta)$ should be compatible with that in the calculation of the values of the function $\varphi(x)$. Moreover, the accuracy in the calculation of a one-dimensional minimum is determined by the number of calculations of the values of $\varphi(x)$ along the direction $-s_k$, i.e. by the number whose choice is related to the choice of the direction of descent. For example, it is undesirable that the determination of s_k require $n+1$

calculations of the values of $\varphi(\mathbf{x})$ while the point of one-dimensional minimum is found from calculations of the values of $\varphi(\mathbf{x})$ whose number differs in an order of magnitude from $n+1$. Although this assertion is not indisputable, it is easy to construct an example with an "ideal" choice of the step which is in no way compatible with the accuracy of the determination of the point of one-dimensional minimum. At any rate, the value of β_k resulting from a one-dimensional minimization procedure must be such that the inequality $\varphi(\mathbf{x}_{k+1}) \leqslant \varphi(\mathbf{x}_k)$ is sure to be fulfilled. We shall write down the basic relation guaranteeing the decrease of the function $\varphi(\mathbf{x})$ that must be satisfied by β_k found by one-dimensional minimization:

$$\varphi(\mathbf{x}_k - \beta_k \mathbf{s}_k) \leqslant (1-\lambda_k)\varphi(\mathbf{x}_k) + \lambda_k \omega_k. \quad (9.12)$$

Here

$$\omega_k = \inf_{\beta \geqslant 0} \varphi(\mathbf{x}_k - \beta \mathbf{s}_k),$$

and the number $\lambda_k \leqslant \frac{\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})}{\varphi(\mathbf{x}_k)-\omega_k} \in (0,1]$ characterizes the calculation accuracy in the determination of the point of one-dimensional minimum of the function $\psi(\beta) = \varphi(\mathbf{x}_k - \beta \mathbf{s}_k)$. From (9.12) it follows that

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant \lambda_k (\varphi(\mathbf{x}_k) - \omega_k).$$

This and (9.7) imply

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \left[1 + \frac{\mu_0}{\eta^2} \sum_{k=0}^{m-1} \frac{\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})}{\| \varphi'(\mathbf{x}_k) \|^2}\right]^{-1}$$

$$\leqslant \mu_0 \left[1 + \frac{\mu_0}{\eta^2} \sum_{k=0}^{m-1} \lambda_k \frac{\varphi(\mathbf{x}_k)-\omega_k}{\| \varphi'(\mathbf{x}_k)^2 \|}\right]^{-1}.$$

These estimates show that as long as the ratio

$$\frac{\varphi(\mathbf{x}_k)-\omega_k}{\| \varphi'(\mathbf{x}_k) \|^2}$$

remains comparatively large the number of calculations of $\varphi(\mathbf{x})$ can be diminished by reducing the accuracy of the one-dimensional minimization, i.e. by decreasing the feasible value of λ_k.

In the course of calculations it often turns out that, despite the fact that a priori estimates indicate the con-

vergence of the method at a rate of, say, $O(1/m)$, with the growth of the number of iterations the rate of convergence either decreases or the process stops to converge. The loss of accuracy in the calculation of the one-dimensional minimum may lie behind this. The above presented estimates show that the unfavorable situation when

$$\frac{\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})}{\|\varphi'(\mathbf{x}_k)\|^2}<\varepsilon$$

can sometimes be improved by increasing the accuracy of one-dimensional minimization, i.e. by increasing the parameter λ_k in a feasible manner so that the inequality

$$\frac{\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})}{\|\varphi'(\mathbf{x}_k)\|^2}\geqslant\varepsilon$$

is fulfilled. The necessity for increasing the accuracy of one-dimensional minimization usually arises in the neighborhood of the point of minimum or when the point $\mathbf{x}_k$ falls in a so-called "ravine", i.e. when the derivative in certain directions nearly vanishes.

9.4.6. Numerically implementable conditions for selecting the parameter β_k without one-dimensional minimization. As $\beta_k>0$ we take $\beta_k=\beta_k^*$ where β_k^* is the greatest of the numbers β satisfying the inequality

$$\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_k-\beta\mathbf{s}_k)\geqslant q\beta\|\mathbf{s}_k\|\frac{\partial\varphi(\mathbf{x}_k)}{\partial\mathbf{s}_k},\quad q>0. \quad (9.13)$$

The condition (9.12) does not allow one to determine the step constructively whereas the inequality (9.13) makes it possible to use one of the elementary methods for calculating β_k, e.g. the duplication method (see Sec. 9.4.4). It should be noted that since the choice of $\mathbf{s}_k$ quarantees local decrease of the function $\varphi(\mathbf{x})$ in the motion from the point $\mathbf{x}_k$ along the direction $-\mathbf{s}_k$, we have

$$\|\mathbf{s}_k\|\frac{\partial\varphi(\mathbf{x}_k)}{\partial\mathbf{s}_k}=\langle\varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle>0.$$

Let us prove that there is a positive number $\bar{\beta}_k$ satisfying (9.13). Indeed, if we take

$$\bar{\beta}_k=2(1-q)\frac{\langle\varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle}{L\|\mathbf{s}_k\|^2},$$

the application of Lemma 9.2.3 yields

$$\varphi(x_k) - \varphi(x_k - \bar{\beta}_k s_k) \geqslant \bar{\beta}_k \langle \varphi'(x_k), s_k \rangle - \frac{L}{2} \bar{\beta}_k^2 \| s_k \|^2$$

$$= \bar{\beta}_k \left(\langle \varphi'(x_k), s_k \rangle - \frac{L}{2} \bar{\beta}_k \| s_k \|^2 \right) = q \bar{\beta}_k \langle \varphi'(x_k), s_k \rangle$$

$$= q \bar{\beta}_k \| s_k \| \frac{\partial \varphi(x_k)}{\partial s_k}. \qquad (9.14)$$

For $0 < q < 1$ we obtain $\bar{\beta}_k > 0$. Since the problem of finding the maximum number satisfying the inequality (9.13) is solved approximately, we can write down the relation that is satisfied by the parameter β_k determined as the maximum value for a certain $q_k \in (0, 1)$:

$$\varphi(x_k) - \varphi(x_k - \beta_k s_k) \geqslant q_k \beta_k \langle \varphi'(x_k), s_k \rangle. \qquad (9.15)$$

Remark. In Supplement 1 it is proved that the number of duplications required for finding β_k from the condition (9.15) in each iteration is finite.

9.4.7. Denote by α_k the value of the cosine of the angle between the direction of the antigradient $-\varphi'(x_k)$ (i.e. the direction of fastest decrease of the function $\varphi(x)$ at the point x_k) and the direction of descent $-s_k$ at that same point:

$$\alpha_k = \frac{\langle \varphi'(x_k), s_k \rangle}{\| \varphi'(x_k) \| \| s_k \|}. \qquad (9.16)$$

Obviously, $\alpha_k > 0$ since this is a necessary condition for selecting the direction. The theorems on the rate of convergence of unconditional minimization methods will be proved under the assumptions that the sequence $\{x_k\}$ is constructed by the formula (9.11), $\alpha_k > 0$, and β_k satisfies either (9.12) or (9.15).

9.4.8. Lemma. *If the function* $\varphi(x)$ *is bounded from below and belongs to the class* $C^{1,1}(E_n)$ *and the sequence* $\{x_k\}$ *is constructed by the formula* (9.11) *and satisfies the condition* (9.12), *then the inequality*

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{1}{2L} \lambda_k \alpha_k^2 \| \varphi'(x_k) \|^2$$

is fulfilled.

Proof. It follows from (9.12) that for any values of the parameter $\beta \geqslant 0$ we have

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k (\varphi(x_k) - \omega)$$

$$\geqslant \lambda_k (\varphi(x_k) - \varphi(x_k - \beta s_k)).$$

Whence, using Lemma 9.2.3 and the definition of the parameter α_k, we derive the inequality

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k \left(\beta \langle \varphi'(x_k), s_k \rangle - \frac{1}{2} L\beta^2 \| s_k \|^2\right)$$

$$= \lambda_k \left(\beta\alpha_k \| \varphi'(x_k) \| \| s_k \| - \frac{1}{2} L\beta^2 \| s_k \|^2\right)$$

holding for any value of β. The right-hand side of the inequality is a quadratic function of β and attains its maximum at

$$\beta = \frac{\alpha_k}{L} \frac{\| \varphi'(x_k) \|}{\| s_k \|}.$$

Thus, substituting this value we find

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{1}{2L} \lambda_k \alpha_k^2 \| \varphi'(x_k) \|^2. \quad \triangle$$

9.4.9. Lemma. *If the function* $\varphi(x)$ *is bounded from below and belongs to the class* $C^{1,1}(E_n)$ *and the sequence* $\{x_k\}$ *is constructed according to* (9.11) *and satisfies the condition* (9.15) *then the inequality*

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{1}{L} \gamma_k \alpha_k^2 \| \varphi'(x_k) \|^2, \quad \gamma_k = 2q_k(1 - q_k)$$

takes place.

Proof. Since β_k is the maximum number satisfying the inequality (9.15) and, as was shown (see (9.14)), the parameter

$$\bar{\beta}_k = 2(1 - q_k) \frac{\langle \varphi'(x_k), s_k \rangle}{L \| s_k \|^2}$$

satisfies (9.15), we have $\beta_k \geqslant \bar{\beta}_k$, and, in view of the definition of α_k,

$$\varphi(x_k) - \varphi(x_{k+1}) = \varphi(x_k) - \varphi(x_k - \beta_k s_k)$$

$$\geqslant q_k \beta_k \langle \varphi'(x_k), s_k \rangle \geqslant q_k \bar{\beta}_k \langle \varphi'(x_k), s_k \rangle$$

$$= 2q_k(1 - q_k) \frac{1}{L} \frac{\langle \varphi'(x_k), s_k \rangle^2}{\| s_k \|^2} = \frac{1}{L} \gamma_k \alpha_k^2 \| \varphi'(x_k) \|^2. \quad \triangle$$

9.4.10. Below we shall prove convergence theorems for relaxational minimization processes. Before proceeding to formal proofs some ideas should be formulated. For smooth non-convex functions we shall prove that rela-

xation sequences converge to the set $\{x^* \in X: \varphi'(x^*) = 0\}$. Since $\{\varphi(x_k)\}$ is a monotonically decreasing sequence, the relaxation process never leads us to a point of local maximum (it is natural to assume that $\varphi'(x_0) \neq 0$). Further, it is unlikely that the process leads to saddle points, which is confirmed by the experience gained in the application of relaxational minimization methods. Hence, in real situations we have convergence to the set of local minima and, in the majority of cases, to a point of local minimum. The more so, since based on some "physical" considerations, when solving specific applied problems we usually know a certain neighborhood of the desired point x^* where the initial point x_0 is taken. Although we cannot exclude the above-mentioned points from our formal argument, what has been said should be taken into consideration when analyzing the convergence theorems.

9.4.11. The set of stationary points. In the case of unconditional minimization problems for smooth functions the set

$$X^* = \{x^*: \varphi'(x^*) = 0\}$$

will be called the *set of stationary points*. We also introduce the sets $X_0 = \{x: \varphi(x) \leqslant \varphi(x_0)\}$ and $X_2^* = X^* \cap X_0$.

9.4.12. A condition guaranteeing convergence. Given a number $\varepsilon > 0$, we construct the set

$$U_\varepsilon = \{x: \rho(x, X_0^*) \leqslant \varepsilon\},$$

where

$$\rho(x, X_0^*) = \inf_{x^* \in X_0^*} \| x - x^* \|^2.$$

The condition guaranteeing the convergence of a relaxation sequence consists in the following: *for each* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that* $\| \varphi'(x_k) \| \geqslant \delta$ *for all* $x_k \in X_0^* \setminus U_\varepsilon$.

9.4.13. For functions bounded from below on the set X_0^*, and it is these functions for which the convergence will be investigated, it is easy to construct an example where the absence of the condition 9.4.12 results in a violation of convergence of a relaxation sequence to the set X^* (see Fig. 9.2).

The condition 9.4.12 is fulfilled for a sufficiently broad class of functions. For instance, when $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$

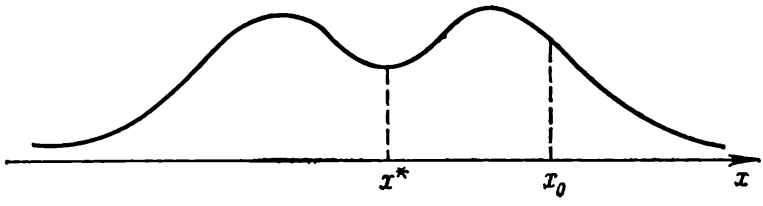


Fig. 9.2

and the set X_0 is bounded, the validity of the condition 9.4.12 follows from Theorem 9.2.7.

9.4.14. Assumptions. We shall assume that the function $\varphi(\mathbf{x})$ satisfies the following conditions:

(1) $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$,
(2) $X_0^* \neq \varnothing$,
(3) $\varphi(\mathbf{x})$ is bounded from below on the set X_0.

The sequence $\{\mathbf{x}_k\}$ will be assumed to possess the following properties:

(4) $\mathbf{x}_0$ is an arbitrary point,
(5) $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k$, $k = 0, 1, \ldots,$ (9.11)
(6) $\alpha_k \geqslant \alpha > 0$, $k = 0, 1, \ldots,$
(7) either the condition

$$\varphi(\mathbf{x}_k - \beta_k \mathbf{s}_k) \leqslant (1 - \lambda_k)\,\varphi(\mathbf{x}_k) + \lambda_k \omega_k \quad (9.12)$$
$$\text{for } 0 < \lambda \leqslant \lambda_k \leqslant 1,\ k = 0, 1, \ldots,$$

or the condition

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta_k \mathbf{s}_k) \geqslant q_k \beta_k \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle \quad (9.15)$$
$$\text{for } 0 < \varepsilon \leqslant q_k \leqslant 1 - \varepsilon,\ k = 0, 1, \ldots,\ \varepsilon < 1/2,$$

is fulfilled.

9.4.15. Theorem. *If the assumptions* 9.4.14 *and the condition* 9.4.12 *are fulfilled, then*

$$\lim_{k \to \infty} \rho(\mathbf{x}_k, X_0^*) = 0. \quad (9.17)$$

Proof. The assumption (7) and Lemmas 9.4.8 and 9.4.9 imply that $0 < \gamma \leqslant 2\varepsilon^2 \leqslant \gamma_k \leqslant 1/2$ and

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant C \| \varphi'(\mathbf{x}_k) \|^2$$

for $C = \frac{1}{L}\alpha^2 \min\left\{\frac{1}{2}\lambda, \gamma\right\}$. The sequence $\{\varphi(\mathbf{x}_k)\}$ is monotonic and bounded from below, and therefore

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \to 0 \text{ for } k \to \infty,$$

$$f(\mathbf{x}_k) \equiv \| \varphi'(\mathbf{x}_k) \|^2 \to 0 \text{ for } k \to \infty.$$

The validity of (9.17) readily follows from Theorem 9.2.5 for $X \equiv X_0$ and $Y \equiv X_0^*$. $\triangle$

9.4.16. Discussion. The above theorem indicates the important role of the choice of the initial point $\mathbf{x}_0$. If this point is appropriately chosen, e.g. so that the set X_0 contains a single stationary point, i.e. the point of global minimum of the function $\varphi(\mathbf{x})$, then (9.17) implies that $\mathbf{x}_k \to \mathbf{x}^*$ for $k \to \infty$. It is obvious that for convex functions the sequence $\{\mathbf{x}_k\}$ converges to the set of points of global minimum and that for strongly convex functions it converges to the single point $\mathbf{x}^*$. In both cases the existence of convergence does not depend on the choice of the point $\mathbf{x}_0$ although its inappropriate choice may substantially decelerate the whole minimization process.

9.4.17. On the violation of relaxation. In practice, during the minimization process there often arises a situation where relaxation is violated, i.e. the sequence $\{\varphi(\mathbf{x}_k)\}$ fails to decrease monotonically. Such violations are sometimes introduced intentionally in order to avoid the deceleration of the minimization process. Usually in such cases use is made of certain heuristic techniques making it possible to restore relaxation after a finite number of steps. It is obvious that here the convergence theorem is true as well, however not for the whole sequence but for its part that remains after all the points where relaxation is violated have been deleted.

9.4.18. We now proceed to the derivation of estimates for the rate of convergence of descent methods of solving unconditional minimization problems for convex and strongly convex functions. For convex functions these estimates are constructed under a rather restrictive condition that the set X_0 is bounded. We note that in this case the assumptions (2) and (3) in Sec. 9.4.14 and the condition 9.4.12 hold, which played an important role in the proof of the foregoing theorem.

9.4.19. Theorem. *Assume that*

(1) *the function* $\varphi(\mathbf{x})$ *is convex and belongs to the class* $C^{1,1}(E_n)$,

(2) $\operatorname{diam} X_0 = \eta < \infty$,

(3) *the sequence* $\{\mathbf{x}_k\}$ *is constructed according to the formula* (9.11),

(4) $\alpha_k > 0$, $k = 0, 1, \ldots$.

If the condition (9.12) *is fulfilled, then*

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \left[1 + C\mu_0 \sum_{k=0}^{m-1} \lambda_k \alpha_k^2\right]^{-1}, \quad m = 1, 2, \ldots, \tag{9.18}$$

for any $0 < C \leqslant \frac{1}{2L\eta^2}$.

In case the condition (9.15) *holds, we have*

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \left[1 + C\mu_0 \sum_{k=0}^{m-1} \gamma_k \alpha_k^2\right]^{-1}, \quad m = 1, 2, \ldots, \tag{9.19}$$

for any $0 < C \leqslant \frac{1}{L\eta^2}$.

The proof obviously follows from Lemmas 9.4.8 and 9.4.9 and Theorem 9.3.3. $\triangle$

9.4.20. Remark. The condition (2) in Theorem 9.4.19 can be replaced by the conditions (3) and (4) in Theorem 9.3.4: $X^* \neq \varnothing$ and for any $\varepsilon > 0$ there is $\delta = \delta(\varepsilon) > 0$ such that $\| \varphi'(\mathbf{x}) \| \geqslant \delta$ for all $\mathbf{x}$ such that $\rho(\mathbf{x}, X^*) \geqslant \varepsilon$. In this case Theorem 9.4.19 remains true. Its proof is based on Lemmas 9.4.8 and 9.4.9 and Theorem 9.3.4.

9.4.21. Theorem. *Assume that*

(1) *the function* $\varphi(\mathbf{x})$ *is stronly convex and belongs to the class* $C^{1,1}(E_n)$,

(2) *the sequence* $\{\mathbf{x}_k\}$ *is constructed by formulas* (9.11),

(3) $\alpha_k > 0$, $k = 0, 1, \ldots$.

If the condition (9.12) *holds, then*

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \exp\left\{-\frac{\rho}{2L} \sum_{k=0}^{m-1} \lambda_k \alpha_k^2\right\}, \quad m = 1, 2, \ldots, \tag{9.20}$$

$$\| x_m - x^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp \left\{ - \frac{\rho}{2L} \sum_{k=0}^{m-1} \lambda_k \alpha_k^2 \right\}, \quad m = 1, 2, \ldots \quad (9.21)$$

If the condition (9.15) *is fulfilled, then*

$$\varphi(x_m) - \varphi(x^*) \leqslant \mu_0 \exp \left\{ - \frac{\rho}{L} \sum_{k=0}^{m-1} \gamma_k \alpha_k^2 \right\}, \quad m = 1, 2, \ldots, \quad (9.22)$$

$$\| x_m - x^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp \left\{ - \frac{\rho}{L} \sum_{k=0}^{m-1} \gamma_k \alpha_k^2 \right\}, \quad m = 1, 2, \ldots \quad (9.23)$$

The proof is an obvious consequence of Lemmas 9.4.8 and 9.4.9 and Theorem 9.3.5. △

9.4.22. Discussion. It is obvious that in the sense of the inequalities (9.18)-(9.23) the most favorable estimates appear when α_k and λ_k are close to unity and γ_k is close to 1/2. However, this requires a high accuracy in the calculation of the gradient of the function $\varphi(x)$ at the points x_k, in the one-dimensional minimization, and also in the determination of the maximum value of β_k from the condition (9.13), which often reduces the effectiveness of the chosen minimization process.

We now assume that the direction of descent is selected under the condition $\alpha_k \geqslant \alpha > 0$ and the accuracy of the calculation of β_k is such that $0 < \lambda \leqslant \lambda_k < 1$ or $0 < \varepsilon \leqslant q_k \leqslant 1 - \varepsilon$. In this case the conditions (9.18) and (9.19) imply an estimate of the form

$$\varphi(x_m) - \varphi(x^*) < C \frac{1}{m}, \quad m = 1, 2, \ldots,$$

and (9.20)-(9.23) imply

$$\varphi(x_m) - \varphi(x^*) \leqslant \mu_0 \exp\{-Cm\}, \quad m = 1, 2, \ldots,$$

$$\| x_m - x^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp\{-Cm\}, \quad m = 1, 2, \ldots,$$

where C designates constants independent of the index m and the parameter μ_0.

9.4.23. Classification. The choice of the way the sequence $\{x_k\}$ is constructed depends substantially on the properties of the function to be minimized, the information used in each iteration, and also the technical means for performing the computation procedures.

The methods can be classified into three groups: *zero-order methods* where only the values of the minimized function are used; *first-order methods* using, in addition to the above, first derivatives; *second-order methods* which also use second derivatives.

9.5. First- and Second-Order Methods

9.5.1. The gradient descent method. One of the most widespread relaxational minimization methods related to the computation of gradient is the descent method in the direction of the antigradient of the function under minimization. The following reasons can be given in favor of this choice of the direction of descent. Since the antigradient, i.e. $-\varphi'(x_k)$, at the point x_k indicates the direction of the fastest decrease of the function, it seems natural to move from the point x_k in this direction.

The descent method where $s_k = \varphi'(x_k)$ is called the *gradient descent method.*

The parameter β_k is calculated in the gradient descent method in a traditional manner by applying one of the one-dimensional minimization methods to the function $\psi(\beta) = \varphi(x_k - \beta\varphi'(x_k))$.

The gradient method possesses an important property related to the choice of the step by the conditions (9.15). The parameter β_k satisfying (9.15) guarantees strong and monotonic convergence of the sequence $\{x_k\}$, $x_{k+1} = x_k - \beta_k\varphi'(x_k)$, to a point $x^* \in X^*$ determined by the choice of the initial point x_0.

Indeed, let the function $\varphi(x) \in C^{1,1}(E_n)$ be convex, let $X^* \neq \varnothing$, and let y be a fixed element of the set X^*. The convexity of $\varphi(x)$ and the condition (9.15) imply

$$\langle \varphi'(x_k), x_k - y\rangle \geqslant \varphi(x_k) - \varphi(y) \geqslant \varphi(x_k) - \varphi(x_{k+1})$$
$$\geqslant q_k\beta_k \| \varphi'(x_k) \|^2,$$

and therefore

$$-2\beta_k \langle \varphi'(x_k),\ x_k - y\rangle + \beta_k^2 \| \varphi'(x_k)\|^2 \leqslant \beta_k^2 (1-2q_k) \| \varphi'(x_k)\|^2,$$

whence for $1/2 < q_k < 1$ we obtain

$$\| x_{k+1} - y\|^2 = \| x_k - y\|^2 - 2\beta_k \langle \varphi'(x_k),\ x_k - y\rangle + \beta_k^2 \| \varphi'(x_k)\|^2 < \| x_k - y\|^2.$$

Hence, the sequence $\{\| x_k - y\|\}$ is monotonic and bounded. We select from $\{x_k\}$ a convergent subsequence $\{x_{k_j}\}$ and let $x^* = \lim\limits_{j\to\infty} x_{k_j}$. Then $x^* \in X^*$ since $\varphi'(x^*) = 0$ and $\lim\limits_{k\to\infty} \| x_k - x^*\| = 0$ because the sequence $\{\| x_k - x^*\|\}$ is monotonic for any fixed element x^*. It is obvious that the element x^* of the set X^* to which the sequence $\{x_k\}$ converges monotonically is determined by the choice of the initial point x_0. It is also evident that the theorem on estimates 9.4.19 remains valid for this choice of the steplength when the condition diam $X_0 = \eta < \infty$ is replaced by $X^* \neq \varnothing$.

In a special case when X^* is a linear manifold this method results in strong convergence to the projection of the point x_0 on the set X^*. Thus, the choice of β_k guarantees the convergence of the gradient method to the normal solution of the original problem relative to x_0 (see Sec. 7.3.2). For the proof of this property and also the fact that the method possesses a property of self-regularization see Supplement 6.

9.5.2. The generalized gradient descent method. It is often advisable to take as the direction of descent not the antigradient of the function $\varphi(x)$ but

$$-s_k = -A_k \varphi'(x_k),$$

where the A_k are symmetric and positive definite matrices for all $k = 0, 1, \ldots$. There exist various methods of constructing the matrices A_k. Below we present some of them but before this the convergence of the process is investigated.

The sequence $\{x_k\}$ will be constructed by the formula

$$x_{k+1} = x_k - \beta_k A_k \varphi'(x_k), \quad k = 0, 1, \ldots,$$

where the parameters β_k will be chosen so that they satisfy either the condition (9.12) or the condition (9.15).

9.5.3. Convergence. Assume that the norms of the matrices A_k are uniformly bounded from above: $\| A_k \| < \zeta$ $(k = 0, 1, \ldots)$ and all the eigenvalues of the matrices are bounded from below by a number $\nu > 0$. Then for any $\mathbf{x}$ we have

$$\langle A_k \mathbf{x}, \mathbf{x} \rangle \geqslant \nu_k \langle \mathbf{x}, \mathbf{x} \rangle \geqslant \nu \| \mathbf{x} \|^2, \; k = 0, 1, \ldots,$$

where ν_k is the minimum eigenvalue of the matrix A_k. It follows that

$$\alpha_k = \frac{\langle \varphi'(\mathbf{x}_k), A_k \varphi'(\mathbf{x}_k) \rangle}{\| \varphi'(\mathbf{x}_k) \| \, \| A_k \varphi'(\mathbf{x}_k) \|} \geqslant \frac{\nu \| \varphi'(\mathbf{x}_k) \|^2}{\| \varphi'(\mathbf{x}_k) \| \, \| A_k \| \, \| \varphi'(\mathbf{x}_k) \|} \geqslant \frac{\nu}{\zeta} > 0.$$

Theorem 9.4.15 obviously implies the convergence of the process and Theorems 9.4.19 and 9.4.21 provide estimates for the rate of convergence.

9.5.4. The steepest descent method. When $A_k = E$, i.e. A_k is a unit matrix, and the point of one-dimensional minimum of the function $\psi(\beta) = \varphi(\mathbf{x}_k - \beta \mathbf{s}_k)$ is taken as β_k, the relaxation process is called the *steepest descent method*:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \varphi'(\mathbf{x}_k),$$
$$\beta_k = \arg\min \{\psi(\beta) = \varphi(\mathbf{x}_k - \beta \mathbf{s}_k) : \beta \geqslant 0\}.$$

9.5.5. The variable scale method. We consider a simple example. Let

$$\varphi(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i x_i^2, \quad \alpha_i > 0, \; i = \overline{1, n}.$$

When the numbers α_i are substantially different, the level surfaces of the function $\varphi(\mathbf{x})$ are elongated along those coordinate directions $\mathbf{e}_i$ to which small values of α_i correspond. In such cases the function $\varphi(\mathbf{x})$ is usually said to have a ravine shape with slant slopes corresponding to small values of α_i and steep slopes corresponding to large values of α_i. If the point $\mathbf{x}_k$ lies on a steep slope near the "ravine bottom", then the descent along the antigradient often results in a situation when the point $\mathbf{x}_{k+1}$

falls on the opposite slope. Since the antigradient is almost perpendicular to the ravine bottom, the convergence of the minimization process is substantially decelerated. Performing change of variables one can make the level surfaces become nearly spherical, which guarantees a high rate of convergence of gradient methods, e.g. of the steepest descent method. To this end, in our example one should take $\mathbf{y} = A^{-1}\mathbf{x}$ where

$$A = \begin{bmatrix} \sqrt{\alpha_1} & 0 & \dots & 0 \\ 0 & \sqrt{\alpha_2} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & & \sqrt{\alpha_n} \end{bmatrix}.$$

If, originally, the descent was performed according to the formula $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \varphi'(\mathbf{x}_k)$, then, since $\varphi'(\mathbf{y}) = A\varphi'(A\mathbf{y}) = A\varphi'(\mathbf{x})$, it is now carried out by the formula

$$\mathbf{y}_{k+1} = \mathbf{y}_k - \beta_k A \varphi'(\mathbf{y}_k).$$

When $\varphi(\mathbf{x})$ is a smooth function, one puts

$$\alpha_i = \frac{\partial^2 \varphi(\mathbf{x})}{\partial x_i^2},$$

where $\mathbf{x}$ is the point of one-dimensional minimum along the direction $\mathbf{e}_i$, and, although this transformation does not turn the level surfaces into spheres, it reduces their elongation and thus facilitates the application of gradient minimization methods.

The derivatives $\frac{\partial^2 \varphi(\mathbf{x})}{\partial x_i^2}$ are calculated approximately using their finite-difference approximations. The successive approximations for the scale factors α_i are usually found by applying the cyclic coordinate-wise descent method (see Sec. 9.7.8).

Thus, the variable scale method in combination with the descent method along the antigradient can be interpreted as a generalized gradient descent method where the diagonal matrix A of second derivatives is taken as A_k.

9.5.6. The Davidon-Fletcher-Powell method. In this method the iterations are carried out according to the

following scheme:

$$x_{k+1} = x_k - \beta_k A_k \varphi'(x_k),$$

$$A_{k+1} = A_k + \frac{r_k r_k^T}{\langle r_k, q_k \rangle} - \frac{(A_k q_k)(A_k q_k)^T}{\langle A_k q_k, q_k \rangle}, \quad A_0 = E,$$

$$r_k = x_{k+1} - x_k, \quad q_k = \varphi'(x_{k+1}) - \varphi'(x_k),$$

$$\beta_k = \arg\min \{\psi(\beta) = \varphi(x_k - \beta A_k \varphi'(x_k)): \beta \geqslant 0\},$$

$$k = 0, 1, \ldots.$$

The problem of minimizing a strongly convex quadratic function is solved by this method after a finite number of iterations not exceeding the dimension n of the space E_n. Since it is often natural to assume that the behavior of the function $\varphi(x)$ in the neighborhood of the point of minimum resembles that of a strongly convex quadratic function, the application of the Davidon-Fletcher-Powell method as well as other methods possessing the indicated finiteness property turns out to be advantageous at the final stage of a minimization process.

As a rule, when using this method, the matrix A_k is renewed (or, as is sometimes said, reinitialized) after a finite number of iterations, namely for some index k_0 one puts $A_{k_0} = A_0$.

The Davidon-Fletcher-Powell method can be regarded as a version of the method of conjugate directions.

9.5.7. Convergence. We shall assume that the function $\varphi(x)$ of class $C^{1,1}(E_n)$ possesses the following property:

(A) $\langle \varphi'(x) - \varphi'(y), x - y \rangle > 0 \ \forall x, y \in E_n, x \neq y.$

It should be noted that this is a property of strongly convex functions.

It is obvious that the matrices A_k are symmetric (this is readily proved by induction). We shall prove that the matrix A_{k+1} is positive definite under the induction hypothesis that the matrix A_k is symmetric and positive definite.

Take an arbitrary $x \neq 0$ and, in view of the existence of the symmetric matrix $A_k^{1/2}$, denote $y = A_k^{1/2} x$, $z = A_k^{1/2} q_k$. Then

$$\langle A_k x, x \rangle = \langle A_k^{1/2} A_k^{1/2} x, x \rangle = \langle A_k^{1/2} x, A_k^{1/2} x \rangle = \langle y, y \rangle,$$

$$\langle A_k \mathbf{q}_k, \mathbf{q}_k\rangle = \langle A_k^{1/2}\mathbf{q}_k, A_k^{1/2}\mathbf{q}_k\rangle = \langle \mathbf{z}, \mathbf{z}\rangle,$$
$$\langle A_k \mathbf{q}_k, \mathbf{x}\rangle = \langle A_k^{1/2}\mathbf{q}_k, A_k^{1/2}\mathbf{x}\rangle = \langle \mathbf{z}, \mathbf{y}\rangle.$$

Further,

$$\text{(B)}\ \langle A_{k+1}\mathbf{x}, \mathbf{x}\rangle = \langle A_k \mathbf{x}, \mathbf{x}\rangle + \frac{\langle \mathbf{r}_k, \mathbf{x}\rangle^2}{\langle \mathbf{r}_k, \mathbf{q}_k\rangle} - \frac{\langle A_k\mathbf{q}_k, \mathbf{x}\rangle^2}{\langle A_k\mathbf{q}_k, \mathbf{q}_k\rangle}$$
$$= \langle \mathbf{y}, \mathbf{y}\rangle - \frac{\langle \mathbf{z}, \mathbf{y}\rangle^2}{\langle \mathbf{z}, \mathbf{z}\rangle} + \frac{\langle \mathbf{r}_k, \mathbf{x}\rangle^2}{\langle \mathbf{r}_k, \mathbf{q}_k\rangle} = \frac{\langle \mathbf{y}, \mathbf{y}\rangle \langle \mathbf{z}, \mathbf{z}\rangle - \langle \mathbf{z}, \mathbf{y}\rangle^2}{\langle \mathbf{z}, \mathbf{z}\rangle}$$
$$+ \frac{\langle \mathbf{r}_k, \mathbf{x}\rangle^2}{\langle \mathbf{r}_k, \mathbf{q}_k\rangle}.$$

The Cauchy-Bunyakovskiĭ inequality

$$|\langle \mathbf{z}, \mathbf{y}\rangle| \leqslant \|\mathbf{z}\|\,\|\mathbf{y}\|$$

implies that

$$\langle \mathbf{y}, \mathbf{y}\rangle \langle \mathbf{z}, \mathbf{z}\rangle - \langle \mathbf{z}, \mathbf{y}\rangle^2 \geqslant 0,$$

where the sign of equality appears only when $\mathbf{y} = \alpha\mathbf{z}$. However, in this case $A_k^{1/2}\mathbf{x} = \alpha A_k^{1/2}\mathbf{q}_k$, and, since the matrix $A_k^{1/2}$ is non-degenerate, we have $\mathbf{x} = \alpha\mathbf{q}_k$. We note that $\alpha \neq 0$ since $\mathbf{x} \neq 0$. The condition (A) implies $\langle \mathbf{q}_k, \mathbf{r}_k\rangle > 0$, and therefore $\langle \mathbf{r}_k, \mathbf{x}\rangle = \alpha \langle \mathbf{r}_k, \mathbf{q}_k\rangle > 0$. Hence, for the last term on the right-hand side of (B) we have

$$\frac{\langle \mathbf{r}_k, \mathbf{x}\rangle^2}{\langle \mathbf{r}_k, \mathbf{q}_k\rangle} > 0.$$

Finally, (B) implies $\langle A_{k+1}\mathbf{x}, \mathbf{x}\rangle > 0$ for $\mathbf{x} \neq 0$, i.e. the matrix A_{k+1} is positive definite.

When the iteration process is carried out with reinitialization of A_k, it is possible to prove that for a twice differentiable strongly convex function there must be

$$\|A_k\| \leqslant \zeta \quad \text{and} \quad \langle A_k\mathbf{x}, \mathbf{x}\rangle \geqslant \nu\|\mathbf{x}\|^2, \quad \nu > 0,$$
$$k = 0, 1, \ldots .$$

Therefore from Theorem 9.4.21 (also see Sec. 9.5.3) it follows that

$$\|\mathbf{x}_m - \mathbf{x}^*\|^2 \leqslant \frac{2}{\rho}\exp\{-Cm\}.$$

9.5.8. Newton's method. If the functtion $\varphi(\mathbf{x})$ is strongly convex and sufficiently smooth in a neighborhood U of

the point $\mathbf{x}^*$, then for $\mathbf{x}_k \in U$ the function

$$\psi(\mathbf{x}) = \varphi(\mathbf{x}_k) + \langle \varphi'(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle + \frac{1}{2} \langle \varphi''(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle$$

approximates $\varphi(\mathbf{x})$ to a high degree of accuracy. The minimum of the function $\psi(\mathbf{x})$ (we mean the local minimum in the neighborhood U) is attained at the point

$$\mathbf{x} = \mathbf{x}_k - (\varphi''(\mathbf{x}_k))^{-1} \varphi'(\mathbf{x}_k)$$

since $\psi'(\mathbf{x}) = \varphi'(\mathbf{x}_k) + \varphi''(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k)$.

Hence, as a direction of descent it is advisable to take

$$-\mathbf{s}_k = -(\varphi''(\mathbf{x}_k))^{-1} \varphi'(\mathbf{x}_k).$$

In essence, in the gradient descent method we confined ourselves to a linear approximation of the function $\varphi(\mathbf{x})$ and took, as a direction of descent, the direction of decrease of the linear function $\psi(\mathbf{x}) = \varphi(\mathbf{x}_k) + \langle \varphi'(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k \rangle$, i.e. $-\psi'(\mathbf{x}) = -\varphi'(\mathbf{x}_k) = -\mathbf{s}_k$. In Newton's method we resort to the quadratic approximation of the function $\varphi(\mathbf{x})$, which allows us to hope that when the point $\mathbf{x}_k$ lies in a sufficiently small neighborhood of the point $\mathbf{x}^*$, the direction $-\mathbf{s}_k = -(\varphi''(\mathbf{x}_k))^{-1} \varphi'(\mathbf{x}_k)$ provides a higher rate of convergence of the iterative process.

To derive an estimate for the rate of convergence we shall assume additionally that the matrix $\varphi''(\mathbf{x})$ is nondegenerate for all $\mathbf{x} \in U$.

Denote $F(\mathbf{x}) = \varphi'(\mathbf{x})$. In Newton's method to solve the equation $F(\mathbf{x}) = \mathbf{0}$ use is made of a recurrence relation written in the form

$$\mathbf{x}_{k+1} = \mathbf{x}_k - (F'(\mathbf{x}_k))^{-1} F(\mathbf{x}_k).$$

We perform the following transformations:

$$\begin{aligned} \mathbf{x}_{k+1} - \mathbf{x}^* &= \mathbf{x}_k - \mathbf{x}^* - (F'(\mathbf{x}_k))^{-1} F(\mathbf{x}_k) \\ &= (F'(\mathbf{x}_k))^{-1} (F'(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^*) - F(\mathbf{x}_k)) \\ &= (F'(\mathbf{x}_k))^{-1} (F'(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^*) - (F(\mathbf{x}_k) - F(\mathbf{x}^*))). \end{aligned}$$

Since $F'(\mathbf{x})$ and $F''(\mathbf{x})$ are continuous in the neighborhood of the point $\mathbf{x}^*$ and hence are bounded, we have

$$F'(\mathbf{x}_k) = F'(\mathbf{x}^*) + \omega(\mathbf{x}_k, \mathbf{x}^*), \quad \|\omega(\mathbf{x}_k, \mathbf{x}^*)\| \leqslant C_1 \|\mathbf{x}_k - \mathbf{x}^*\|;$$

$$F(\mathbf{x}_k) - F(\mathbf{x}^*) = F'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*) + \varepsilon(\mathbf{x}_k, \mathbf{x}^*), \quad \|\varepsilon(\mathbf{x}_k, \mathbf{x}^*)\| \leqslant C_2 \|\mathbf{x}_k - \mathbf{x}^*\|^2.$$

Therefore

$$\|F(\mathbf{x}_k) - F(\mathbf{x}^*) - F'(\mathbf{x}_k)(\mathbf{x}_k - \mathbf{x}^*)\| = \|F'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*) + \varepsilon(\mathbf{x}_k, \mathbf{x}^*) - F'(\mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*) - \omega(\mathbf{x}_k, \mathbf{x}^*)(\mathbf{x}_k - \mathbf{x}^*)\| \leqslant \|\varepsilon(\mathbf{x}_k, \mathbf{x}^*)\| + \|\omega(\mathbf{x}_k, \mathbf{x}^*)\| \|\mathbf{x}_k - \mathbf{x}^*\| \leqslant C_3 \|\mathbf{x}_k - \mathbf{x}^*\|^2.$$

Finally, we derive

$$\|\mathbf{x}_{k+1} - \mathbf{x}^*\| \leqslant C \|\mathbf{x}_k - \mathbf{x}^*\|^2.$$

9.5.9. A modification of Newton's method. Since the calculation of the Hesse matrix $\varphi''(\mathbf{x}_k)$ in each iteration and its inversion require much computing time, use is made of various modifications of the method. The simplest of them consists in the following. One sets a number $m \geqslant 1$ determining a computation cycle consisting of m iterations. In each lth cycle all the iterations are carried out using the constant matrix $(\varphi''(\mathbf{x}))^{-1}$:

$$\mathbf{x}_{lm+i+1} = \mathbf{x}_{lm+i} - \beta_{lm+i} (\varphi''(\mathbf{x}_{lm}))^{-1} \varphi'(\mathbf{x}_{lm+i}), \quad i = \overline{0, m-1},$$

after which one passes to the next cycle with the matrix $(\varphi''(\mathbf{x}_{l(m+1)}))^{-1}$. This leads to a decrease in the rate of convergence; however, the total machine time in each iteration is substantially reduced.

9.5.10. Accelerated Newton's method. In Newton's method, proceeding from a point $\mathbf{x}$ belonging to a sufficiently small neighborhood U of the point $\mathbf{x}^*$, the point $\mathbf{y} = \mathbf{x} - (\varphi''(\mathbf{x}))^{-1}\varphi'(\mathbf{x})$ is constructed, which satisfies the inequality

$$\|\mathbf{y} - \mathbf{x}^*\| \leqslant C \|\mathbf{x} - \mathbf{x}^*\|^2.$$

Below we consider a method of finding the point $\mathbf{y}_p$ $(p \geqslant 1)$ satisfying the inequality

$$\|\mathbf{y}_p - \mathbf{x}^*\| \leqslant C \|\mathbf{x} - \mathbf{x}^*\|^{p+1}. \tag{9.24}$$

As in Sec. 9.5.8, we shall assume that the function $\varphi(\mathbf{x})$ is strongly convex and sufficiently smooth in a neighborhood U of the point $\mathbf{x}^*$.

We demonstrate the application of the method by an example where $p = 2$. Consider the equation

$$F(\mathbf{x}) = \varphi'(\mathbf{x}) = \mathbf{0}. \tag{9.25}$$

Using Taylor's formula

$$F(\mathbf{y}) = F(\mathbf{x}) + F'(\mathbf{x})[\mathbf{y}-\mathbf{x}] + \frac{1}{2}F''(\mathbf{x})[\mathbf{y}-\mathbf{x}]^2 + \omega(\mathbf{x}, \mathbf{y}),$$

where $\| \omega(\mathbf{x}, \mathbf{y}) \| = o(\| \mathbf{x} - \mathbf{y} \|^2)$, we take as $\mathbf{y}$ an approximate solution of the nonlinear system of equations

$$F(\mathbf{x}) + F'(\mathbf{x})[\mathbf{y}-\mathbf{x}] + \frac{1}{2}F''(\mathbf{x})[\mathbf{y}-\mathbf{x}]^2 = \mathbf{0}. \tag{9.26}$$

To find the element $\mathbf{y}$ we linearize the system (9.26) in the following way: the difference $\mathbf{y} - \mathbf{x}$ in the third term is replaced by the vector $-(F'(\mathbf{x}))^{-1} F(\mathbf{x})$ used in classical Newton's method. This results in a linear system of equations in the unknown $\mathbf{y}$:

$$F(\mathbf{x}) + F'(\mathbf{x})[\mathbf{y}-\mathbf{x}] - \frac{1}{2}F''(\mathbf{x})\{(F'(\mathbf{x})^{-1}F(\mathbf{x})\}[\mathbf{y}-\mathbf{x}] = \mathbf{0}.$$

Its solution is

$$\mathbf{y} = \mathbf{x} - \left\{F'(\mathbf{x}) - \frac{1}{2}F''(\mathbf{x})\{(F'(\mathbf{x}))^{-1}F(\mathbf{x})\}\right\}^{-1} F(\mathbf{x}).$$

Coming back to the function $\varphi(\mathbf{x})$ we derive a computational formula for the next approximation:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \left\{\varphi''(\mathbf{x}_k) - \frac{1}{2}\varphi'''(\mathbf{x}_k)\{(\varphi''(\mathbf{x}_k))^{-1}\varphi'(\mathbf{x}_k)\}\right\}^{-1} \varphi'(\mathbf{x}_k). \tag{9.27}$$

In this case the estimate

$$\| \mathbf{x}_{k+1} - \mathbf{x}^* \| \leqslant C \| \mathbf{x}_k - \mathbf{x}^* \|^3$$

also turns out to hold.

We now pass to the description of the method for the general case $p \geqslant 2$. Consider the matrices $P_k(\mathbf{x})$ and elements $g_k(\mathbf{x})$ satisfying the recurrence relations

$$P_1(\mathbf{x}) = F'(\mathbf{x}), \quad g_1(\mathbf{x}) = (P_1(\mathbf{x}))^{-1} F(\mathbf{x}),$$

$$P_m(\mathbf{x}) = \sum_{k=1}^{m} \frac{1}{k!} F^{(k)}(\mathbf{x})[g_{m-1}(\mathbf{x})]^{k-1},$$

$$g_m(\mathbf{x}) = -(P_m(\mathbf{x}))^{-1} F(\mathbf{x}), \quad m = \overline{1, p}.$$

We shall assume that the derivatives $F^{(k)}(\mathbf{x})$ $(k = \overline{1, m})$ are bounded jointly in the neighborhood U of the point $\mathbf{x}^*$ and the matrix $F'(\mathbf{x}^*)$ is non-degenerate.

Theorem. *For all points* $\mathbf{x}$ *lying sufficiently close to* $\mathbf{x}^*$ *the matrices* $P_k(\mathbf{x})$ *are non-degenerate and the condition*

$$\mathbf{x}^* = \mathbf{x} + g_k(\mathbf{x}) + \omega_k(\mathbf{x}^*, \mathbf{x}) \qquad (9.28)$$

holds, where $\| \omega_k(\mathbf{x}^*, \mathbf{x})\| \leqslant C \| \mathbf{x}^* - \mathbf{x} \|^{k+1}$ $(k = \overline{1, p})$.

Proof. We prove the theorem by induction on the index k. For $k = 1$, due to the non-degeneracy of the matrix $F'(\mathbf{x})$, the condition (9.28) is an estimate for the rate of convergence of Newton's method.

Assume that the matrices $P_k(\mathbf{x})$ are non-degenerate and the condition (9.28) holds for all $k \leqslant m - 1$. Let $k = m \leqslant p$. From Taylor's formula and the definition of $P_m(\mathbf{x})$ if follows that

$$\begin{aligned}
\mathbf{0} = F(\mathbf{x}^*) &= F(\mathbf{x}) + P_1(\mathbf{x})[\mathbf{x}^* - \mathbf{x}] \\
&\quad + \sum_{k=2}^{m} \frac{1}{k!} F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^k + \nu(\mathbf{x}^*, \mathbf{x}) \\
&= F(\mathbf{x}) + P_m(\mathbf{x})[\mathbf{x}^* - \mathbf{x}] \\
&\quad + \sum_{k=2}^{m} \frac{1}{k!} \{F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^k \\
&\quad - F^{(k)}(\mathbf{x})[g_{m-1}(\mathbf{x})]^{k-1}[\mathbf{x}^* - \mathbf{x}]\} + \nu(\mathbf{x}^*, \mathbf{x}), \qquad (9.29)
\end{aligned}$$

where $\| \nu(\mathbf{x}^*, \mathbf{x}) \| \leqslant \frac{1}{(m+1)!} \max\limits_{\xi \in [\mathbf{x}^*, \mathbf{x}]} \|F^{(p+1)}(\xi)\| \, \|\mathbf{x}^* - \mathbf{x}\|^{m+1}$.

By the induction hypothesis,

$$g_{m-1}(\mathbf{x}) = \mathbf{x}^* - \mathbf{x} - \omega_{m-1}(\mathbf{x}^*, \mathbf{x}),$$

$$\| \omega_{m-1}(\mathbf{x}^*, \mathbf{x})\| \leqslant C \| \mathbf{x}^* - \mathbf{x} \|^m,$$

and therefore

$$\begin{aligned}
&F^{(k)}(\mathbf{x})[g_{m-1}(\mathbf{x})]^{k-1}[\mathbf{x}^* - \mathbf{x}] \\
&\quad = \sum_{i=0}^{k-1} C_{k-1}^i F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^{k-i}[\omega_{m-1}(\mathbf{x}^*, \mathbf{x})]^i \\
&\quad = F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^k \\
&\quad + \sum_{i=1}^{k-1} C_{k-1}^i F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^{k-i}[\omega_{m-1}(\mathbf{x}^*, \mathbf{x})]^i.
\end{aligned}$$

Consequently,

$$\left\| \sum_{k=2}^{m} \frac{1}{k!} \{F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^k - F^{(k)}(\mathbf{x})[g_{m-1}(\mathbf{x})]^{k-1}[\mathbf{x}^* - \mathbf{x}]\} \right\|$$

$$\leqslant \left\| \sum_{k=2}^{m} \frac{1}{k!} \sum_{i=1}^{k-1} C_{k-1}^{i} F^{(k)}(\mathbf{x})[\mathbf{x}^* - \mathbf{x}]^{k-i}[\omega_{m-1}(\mathbf{x}^*, \mathbf{x})]^i \right\|. \tag{9.30}$$

Since, according to the statement of the theorem, the derivatives $F^{(k)}(\mathbf{x})$ are continuous in the neighborhood of the point $\mathbf{x}^*$, we have

$$\|F^{(k)}(\mathbf{x})[\mathbf{z}]^{k-p}[\mathbf{v}]^p\| \leqslant N\|\mathbf{z}\|^{k-p} \|\mathbf{v}\|^p, \quad N > 0,$$

for any $\mathbf{z}$, $\mathbf{v}$ and $p = \overline{0, k}$. Therefore the right-hand side of (9.30) does not exceed $C \|\mathbf{x}^* - \mathbf{x}\|^{m+1}$ where $C > 0$ is a finite constant.

Thus, (9.29) is written as

$$F(\mathbf{x}) + P_m(\mathbf{x})[\mathbf{x}^* - \mathbf{x}] + \delta(\mathbf{x}^*, \mathbf{x}) = \mathbf{0},$$

where $\|\delta(\mathbf{x}^*, \mathbf{x})\| \leqslant C \|\mathbf{x}^* - \mathbf{x}\|^{m+1}$; moreover, since $P_m(\mathbf{x}^*) = F'(\mathbf{x}^*)$ and, by the induction hypothesis, $\|g_{m-1}(\mathbf{x})\| = \|\mathbf{x}^* - \mathbf{x}\| + o(\|\mathbf{x}^* - \mathbf{x}\|^{m-1})$, the definition of $P_m(\mathbf{x})$ implies that $P_m(\mathbf{x})$ coincides with $F'(\mathbf{x}^*)$ to within the order of magnitude of $\|\mathbf{x}^* - \mathbf{x}\|$. Consequently, the matrices $P_m(\mathbf{x})$ are nondegenerate for all $\mathbf{x}$ lying sufficiently close to $\mathbf{x}^*$, and the condition (9.29) implies

$$\mathbf{x}^* - \mathbf{x} = g_m(\mathbf{x}) - P_m^{-1}(\mathbf{x})\,\delta(\mathbf{x}^*, \mathbf{x}),$$

which proves the assertion. Putting $\mathbf{y}_p = \mathbf{x} + g_p(\mathbf{x})$ we obtain the inequality $\|\mathbf{y}_p - \mathbf{x}^*\| \leqslant C \|\mathbf{x} - \mathbf{x}^*\|^{p+1}$.

We note that for $p = 1$ the method of constructing the point $\mathbf{y}_p$ coincides with ordinary Newton's method.

For $p > 1$ the above method sometimes possesses certain advantages over ordinary Newton's method. In particular, when the derivatives $F^{(k)}(\mathbf{x})$ have a small number of nonzero elements, the amount of calculations in the method is of the same order as that in Newton's method. Moreover, there are problems where the higher derivatives of the function $F(\mathbf{x})$ are calculated much easier than the values of $F(\mathbf{x})$ and $F'(\mathbf{x})$. For instance, this is the

case in the traditional algebraic problem of finding the roots of higher-degree polynomials.

9.5.11. An apparent advantage of Newton's method is its high rate of convergence as compared to other iteration methods. However, the computation time for an iteration in Newton's method (the passage from $\mathbf{x}_k$ to $\mathbf{x}_{k+1}$) is related to the computation and inversion of the matrix of second derivatives $\varphi''(\mathbf{x}_k)$ and may turn out to be so long that this will not be compensated by the determination of $\varphi(\mathbf{x}_{k+1})$ with the help of several iterations of some other less laborious method. Therefore the advisability of application of Newton's method is inseparable from the effectiveness of the whole iterative process, i.e. the total amount of calculations required for solving the problem with a prescribed accuracy. It is these well-known factors that have led to the necessity of elaborating methods which are less time consuming and possess a rate of convergence close to that of Newton's method. That is why methods of quasi-Newton type have appeared. When Newton's method is applied to minimize a strongly convex quadratic function, the problem is solved after one iteration. In methods of quasi-Newton type the matrices A_k are constructed according to recurrence relations so that the successive approximations $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k A_k \varphi'(\mathbf{x}_k)$ minimize a strongly convex quadratic function after a finite number of iterations. The method with variable metric presented in Sec. 9.5.6 is a typical representative of quasi-Newton type methods.

In the next section we consider methods which also prove finite in application to quadratic minimization problems, and from this standpoint they should be related to methods of quasi-Newton type.

9.6. Method of Conjugate Directions

9.6.1. The scheme of the method.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k, \qquad k = 0, 1, \ldots, \tag{9.11}$$

$$\mathbf{s}_0 = \varphi'(\mathbf{x}_0), \quad \mathbf{s}_k = \varphi'(\mathbf{x}_k) - \xi_k \mathbf{s}_{k-1}, \qquad k = 1, 2, \ldots, \tag{9.31}$$

$$\beta_k = \arg\min \{\varphi(\mathbf{x}_k - \beta \mathbf{s}_k): \beta \geqslant 0\}, \qquad k = 0, 1, \ldots. \tag{9.32}$$

Various modifications of the method of conjugate directions differ in the ways the parameter ξ_k is chosen. We

note that for $\xi_k = 0$ this method degenerates into the method of steepest descent.

The condition (9.25) for the choice of the parameter β_k determines the following two specific properties of the sequence $\{x_k\}$.

9.6.2. Lemma. *The sequence* $\{x_k\}$ *constructed for a differentiable function* $\varphi(x)$ *according to the scheme* (9.11), (9.31), (9.32) *satisfies the relations*

$$\langle \varphi'(x_{k+1}), s_k \rangle = 0, \quad k = 0, 1, \ldots, \tag{9.33}$$

$$\langle \varphi'(x_k), s_k \rangle = \| \varphi'(x_k) \|^2, \quad k = 0, 1, \ldots \tag{9.34}$$

Proof. It follows from (9.32) that for $\beta_k > 0$ and $\beta_k = 0$ we have

$$\frac{d}{d\beta} \varphi(x_k - \beta s_k)|_{\beta=\beta_k} = 0,$$

and

$$\frac{d}{d\beta} \varphi(x_k - \beta s_k)|_{\beta=0} \geqslant 0,$$

respectively.

If $\beta_k > 0$, then

$$0 = \frac{d}{d\beta} \varphi(x_k - \beta s_k)|_{\beta=\beta_k} = -\langle \varphi'(x_k - \beta_k s_k), s_k \rangle = -\langle \varphi'(x_{k+1}), s_k \rangle.$$

The fact that (9.33) holds for $\beta_k = 0$ as well will be proved by induction. If $\beta_0 = 0$, then $x_1 = x_0$ and $s_0 = \varphi'(x_0)$ imply

$$0 \leqslant \frac{d}{d\beta} \varphi(x_0 - \beta s_0)|_{\beta=0} = -\langle \varphi'(x_1), s_0 \rangle = -\| \varphi'(x_0) \|^2,$$

whence

$$\langle \varphi'(x_1), s_0 \rangle = 0.$$

Let the relation $\langle \varphi'(x_k), s_{k-1} \rangle = 0$ hold. We shall prove that

$$\langle \varphi'(x_{k+1}), s_k \rangle = 0$$

for $\beta_k = 0$. Since $x_{k+1} = x_k$, we obtain from (9.31) the relation

$$0 \leqslant \frac{d}{d\beta} \varphi(x_k - \beta s_k)|_{\beta=0} = -\langle \varphi'(x_{k+1}), s_k \rangle = -\langle \varphi'(x_k), s_k \rangle = -\langle \varphi'(x_k), \varphi'(x_k) - \xi_k s_{k-1} \rangle = -\| \varphi'(x_k) \|^2,$$

which implies (9.33).

Finally, (9.34) is an obvious consequence of (9.31) and (9.33). △

9.6.3. Lemma. *If*

$$|\xi_k| \leqslant C \frac{\| \varphi'(x_k) \|}{\| s_{k-1} \|}, \quad k = 1, 2, \ldots, \tag{9.35}$$

for some $C > 0$, *then*

$$\alpha_k = \frac{\langle \varphi'(x_k), s_k \rangle}{\| \varphi'(x_k) \| \| s_k \|} \geqslant \frac{1}{1+C}. \tag{9.36}$$

The proof is an obvious consequence of (9.31) and (9.35). △

9.6.4. Convergence. It is obvious that, by Theorem 9.4.15, the iterative process is convergent, and Theorems 9.4.19 and 9.4.21 provide an estimate for the rate of convergence.

As examples of realization of the method of conjugate directions we consider two methods of selecting the parameters ξ_k which guarantee the finiteness of the method for the case of quadratic functions.

9.6.5. First method.

$$\xi_k = -\frac{1}{\| \varphi'(x_{k-1}) \|^2} \langle \varphi'(x_k), \varphi'(x_k) - \varphi'(x_{k-1}) \rangle.$$

We shall prove that for a strongly convex function $\varphi(x)$ of class $C^{1,1}(E_n)$ the condition (9.35) is fulfilled. From the assumption of strong convexity it follows that (see Sec. 2.2.5)

$$\langle \varphi'(x_k) - \varphi'(x_{k-1}), x_k - x_{k-1} \rangle \geqslant \rho \| x_k - x_{k-1} \|^2,$$

which, in view of (9.11), (9.33), and (9.34), implies the inequality

$$\begin{aligned} \rho \beta_{k-1}^2 \| s_{k-1} \|^2 &\leqslant -\beta_{k-1} \langle \varphi'(x_k) - \varphi'(x_{k-1}), s_{k-1} \rangle \\ &= \beta_{k-1} \langle \varphi'(x_{k-1}), s_{k-1} \rangle = \beta_{k-1} \| \varphi'(x_{k-1}) \|^2. \end{aligned}$$

Since it is always assumed that $\varphi'(x_k) \neq 0$ for all $k = 0, 1, \ldots$, the proof of Lemma 9.6.2 implies that $\beta_k > 0$ since $\beta_{k-1} = 0$ implies $\| \varphi'(x_{k-1}) \| = 0$. Thus,

$$\| \varphi'(x_{k-1}) \|^2 \geqslant \rho \beta_{k-1} \| s_{k-1} \|^2.$$

Taking this and the condition

$$\| \varphi'(x_k) - \varphi'(x_{k-1}) \| \leqslant L \| x_k - x_{k-1} \|$$

into account we estimate the parameter ξ_k:

$$|\xi_k| = \frac{1}{\| \varphi'(x_{k-1}) \|^2} |\langle \varphi'(x_k), \varphi'(x_k) - \varphi'(x_{k-1}) \rangle|$$
$$\leqslant \frac{1}{\| \varphi'(x_{k-1}) \|^2} \| \varphi'(x_k) \| \| \varphi'(x_k) - \varphi'(x_{k-1}) \|$$
$$\leqslant \frac{1}{\| \varphi'(x_{k-1}) \|^2} \| \varphi'(x_k) \| L \| x_k - x_{k-1} \|$$
$$= \frac{L}{\| \varphi'(x_{k-1}) \|^2} \| \varphi'(x_k) \| \beta_{k-1} \| s_{k-1} \|$$
$$= \frac{L}{\rho \beta_{k-1} \| s_{k-1} \|^2} \| \varphi'(x_k) \| \beta_{k-1} \| s_{k-1} \| = \frac{L}{\rho} \frac{\| \varphi'(x_k) \|}{\| s_{k-1} \|}.$$

Thus, the condition (9.35) is fulfilled, and therefore 9.4.21 implies the estimates

$$\varphi(x_m) - \varphi(x^*) \leqslant \mu_0 \exp\{-C_1 m\},$$
$$\| x_m - x^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp\{-C_1 m\},$$

where $0 < C_1 \leqslant \frac{\rho^3}{2L(\rho + L)^2}$.

9.6.6. Second method.

$$\xi_k = \frac{\langle \varphi''(x_k) s_{k-1}, \varphi'(x_k) \rangle}{\langle \varphi''(x_k) s_{k-1}, s_{k-1} \rangle}.$$

This method of selecting the parameter ξ_k makes it necessary to calculate the matrix of second derivatives at each iteration; however, in contrast to Newton's method, it does not require the inversion of the matrix. It should be noted that the number of multiplications involved in the inversion of an $n \times n$ matrix is proportional to n^3.

An elementary analysis of convergence is useful from the methodical viewpoint since analogous techniques apply to estimates for the rate of convergence of quasi-Newton type methods whose numerous versions can be found in various papers and continue to appear in new mathematical publications.

Assume that a strongly convex function $\varphi(x)$ is twice continuously differentiable and the norm of the matrix $\varphi''(x)$ is uniformly bounded on the bounded set

$X_0 = \{\mathbf{x}: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$:

$$\| \varphi''(\mathbf{x}) \| \leqslant \nu < \infty.$$

Since

$$\langle \varphi'(\mathbf{x} + \varepsilon \mathbf{y}) - \varphi'(\mathbf{x}), \mathbf{y} \rangle = \varepsilon \langle \varphi''(\mathbf{x} + \varepsilon\theta\mathbf{y})\, \mathbf{y}, \mathbf{y} \rangle, \quad 0 \leqslant \theta \leqslant 1,$$

taking into consideration the property of strong convexity (2.5.5) we conclude that $\langle \varphi''(\mathbf{x} + \varepsilon\theta\mathbf{y})\, \mathbf{y}, \mathbf{y} \rangle \geqslant \rho \| \mathbf{y} \|^2$. Passing to the limit in this inequality for $\varepsilon \to 0$ we arrive at the relation

$$\langle \varphi''(\mathbf{x})\, \mathbf{y}, \mathbf{y} \rangle \geqslant \rho \| \mathbf{y} \|^2 \tag{9.37}$$

holding for any $\mathbf{y} \in E_n$. Consequently, the inequality

$$|\xi_k| = \frac{|\langle \varphi''(\mathbf{x}_k)\, \mathbf{s}_{k-1}, \varphi'(\mathbf{x}_k) \rangle|}{|\langle \varphi''(\mathbf{x}_k)\, \mathbf{s}_{k-1}, \mathbf{s}_{k-1} \rangle|} \leqslant \frac{\| \varphi''(\mathbf{x}_k) \| \, \| \mathbf{s}_{k-1} \| \, \| \varphi'(\mathbf{x}_k) \|}{\rho \| \mathbf{s}_{k-1} \|^2} \leqslant \frac{\nu}{\rho} \frac{\| \varphi'(\mathbf{x}_k) \|}{\| \mathbf{s}_{k-1} \|}$$

takes place, and therefore Theorem 9.4.21 implies the estimates

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \exp\{-C_2 m\},$$

$$\| \mathbf{x}_m - \mathbf{x}^* \|^2 \leqslant \frac{2}{\rho} \mu_0 \exp\{-C_2 m\},$$

where

$$0 < C_2 \leqslant \frac{\rho^3}{2L(\rho + \nu)^2}.$$

9.6.7. Discussion. Although the presented estimates are valid for all $m = 1, 2, \ldots$ and guarantee a definite rate of convergence of the conjugate directions method, they, as was already said, do not reveal the following advantage of the method: for a strongly convex quadratic function $\varphi(\mathbf{x})$ each of the presented methods of determining the parameter ξ_k makes it possible to solve the problem after a finite number of steps not exceeding the dimension n of the space E_n. This fact plays an important part when the method is used for solving problems of minimizing non-quadratic functions. Namely, n iterations are performed according to the method of conjugate directions, after which, the method, as we say, is renewed

by putting $\xi_n = 0$, i.e. gradient descent is carried out. Thus, in this case the scheme of the method is

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k, \quad \mathbf{s}_k = \varphi'(\mathbf{x}_k) - \bar{\xi}_k \mathbf{s}_{k-1},$$
$$\bar{\xi}_k = \begin{cases} \xi_k, & k \in I_1, \\ 0, & k \in I_2 = \{0, n, 2n, 3n, \ldots\}, \end{cases}$$
$$I_1 \cup I_2 = \{0, 1, \ldots\}, \quad \varphi(\mathbf{x}_{k+1}) = \min_{\beta \geqslant 0} \varphi(\mathbf{x}_k - \beta \mathbf{s}_k),$$

and the parameter ξ_k is determined based on the finiteness of the method for quadratic functions, for instance, according to one of the above methods. This is the so-called *n-step scheme*.

Finally, it should be noted that the method of conjugate directions is highly sensitive to errors appearing in the computation process since for $\lambda_k \neq 1$ the property of orthogonality (9.33) is violated in (9.12) and therefore the method is no longer finite in application to quadratic functions.

9.6.8. On conjugate directions. Consider a symmetric positive definite matrix A. Vectors $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}$ are said to be *conjugate relative to the matrix* A if

$$\langle \mathbf{s}_i, A\mathbf{s}_j \rangle = 0, \quad i \neq j.$$

It can be proved that for the problem of minimizing the quadratic function $\varphi(\mathbf{x}) = \langle A\mathbf{x}, \mathbf{x} \rangle + \langle \mathbf{b}, \mathbf{x} \rangle$ with a symmetric positive definite matrix A the vectors $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}$ constructed by the method of conjugate directions are conjugate relative to the matrix A. The same relates to the vectors $\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}$ introduced in the method with variable metric, and in this sense it can be regarded as a method of conjugate directions.

9.7. Zero-Order Methods

9.7.1. Coordinate-wise descent. It seems that the simplest deterministic methods for determining the direction of descent is the choice of one of the coordinate vectors $\pm\mathbf{e}_1, \pm\mathbf{e}_2, \ldots, \pm\mathbf{e}_n$ as $\mathbf{s}_k$, owing to which only one of the components of the vector $\mathbf{x}_k$ is changed in each iteration.

There are numerous versions of the coordinate-wise descent method, and only some of them will be considered in this section. We first of all note that in any of these methods one of the two directions $+\mathbf{e}_j$ and $-\mathbf{e}_j$ is taken as $-\mathbf{s}_k$, namely the one to which the inequality

$$\langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle > 0$$

corresponds. When $\frac{\partial \varphi(\mathbf{x}_k)}{\partial x_j} = 0$, one puts $\mathbf{x}_{k+1} = \mathbf{x}_k$ and passes to the next iteration. For the sake of simplicity of the presentation, this case is excluded from consideration since it is irrelevant to the essence of the method, and, the more so, we assume that $\varphi'(\mathbf{x}_k) \neq \mathbf{0}$ $(k = 1, 2, \ldots)$.

9.7.2. Coordinate-wise descent with duplication of step (the first version). We describe the first cycle of the method. It consists of n iterations.

At an arbitrary point $\mathbf{x}_0$ one takes $\mathbf{s}_0 = \pm\mathbf{e}_1$ and determines the parameter β_0 by the duplication method (see 9.4.4) so that $\varphi(\mathbf{x}_1) = \varphi(\mathbf{x}_0 - \beta_0\mathbf{s}_0) < \varphi(\mathbf{x}_0)$. Then $\mathbf{s}_1 = \pm\mathbf{e}_2$ is taken and, putting $\beta = \beta_0$, the parameter β_1 is calculated by means of duplication, etc. Here in each iteration one tries to determine the steplength by the duplication method with minimum number of calculations of the values of the function $\varphi(\mathbf{x})$. The cycle is ended for $k = n - 1$, after which the next cycle is begun by putting $\mathbf{s}_n = \pm\mathbf{e}_1$, etc.

9.7.3. Coordinate-wise descent with duplication of step (the second version). In each cycle of the method the step remains constant during all n iterations in the cycle. The cycle consists in calculating the points $\mathbf{x}_k$, $\mathbf{x}_{k+1}$, . . ., $\mathbf{x}_{k+n-1}$. It is assumed that after the foregoing cycle is completed the parameter $\beta = \beta_{k-1}$ is found.

The $(i+1)$th iteration in a cycle $(0 \leqslant i \leqslant n-1)$:

If

(A) $$\varphi(\mathbf{x}_{k+i} - \beta\mathbf{e}_{i+1}) < \varphi(\mathbf{x}_{k+i}),$$

one puts $\mathbf{x}_{k+i-1} = \mathbf{x}_{k+i} - \beta\mathbf{e}_{i+1}$ and passes to the next iteration.

If

(B) $$\varphi(\mathbf{x}_{k+i} - \beta\mathbf{e}_{i+1}) \geqslant \varphi(\mathbf{x}_{k+i}),$$

then $\varphi(\mathbf{x}_{k+i} + \beta\mathbf{e}_{i+1})$ is calculated.

If

(C) $$\varphi(\mathbf{x}_{k+i}+\beta\mathbf{e}_{i+1})<\varphi(\mathbf{x}_{k+i}),$$

one puts $\mathbf{x}_{k+i+1}=\mathbf{x}_{k+i}+\beta\mathbf{e}_{i+1}$ and passes to the next iteration.

In case

(D) $$\varphi(\mathbf{x}_{k+i}+\beta\mathbf{e}_{i+1})\geqslant\varphi(\mathbf{x}_{k+i}),$$

one puts $\mathbf{x}_{k+i+1}=\mathbf{x}_{k+i}$ and passes to the next iteration.

When the inequalities (B) and (D) hold for all $0\leqslant i\leqslant n-1$, one diminishes the parameter β by putting, as a rule, $\beta=\beta_{k-1}/2$ and passes to the next cycle, i.e. all the procedures of the foregoing cycle are repeated with a twice decreased steplength.

9.7.4. Assumptions. When investigating the convergence of the method 9.7.3 we shall assume that

(1) $\varphi(\mathbf{x})\in C^{1,1}(E_n)$,

(2) X_0 is a bounded set.

9.7.5. Convergence. We first *prove that* $\beta_k\to 0$ *for* $k\to\infty$. If otherwise, there is an index k_0 such that $\beta_k=\beta>0$ for all $k\geqslant k_0$, and, consequently, $\varphi(\mathbf{x}_{k+1})<\varphi(\mathbf{x}_k)\ \forall k\geqslant k_0$. Since $\|\mathbf{x}_{k+1}-\mathbf{x}_k\|=\beta>0\ \forall k\geqslant k_0$, the number of the points falling in the bounded set X_0 is finite, and there is an index $k_1>k_0$ beginning with which there arises a situation where $\mathbf{x}_k\notin X_0\forall k\geqslant k_1$ takes place, which contradicts the definition of the set X_0. $\triangle$

Let us consider all the cycles in which the inequalities (B) and (D) hold simultaneously for all $0\leqslant i\leqslant n-1$. Since the steplength tends to zero for $k\to\infty$, there are an infinite number of such cycles. We index the cycles so that

$$\varphi(\mathbf{x}_m+\beta_m\mathbf{e}_{i+1})\geqslant\varphi(\mathbf{x}_m),\quad i=\overline{0,\ n-1},\quad m=0,\ 1,\ \ldots,$$

$$\varphi(\mathbf{x}_m-\beta_m\mathbf{e}_{i+1})\geqslant\varphi(\mathbf{x}_m),\quad i=\overline{0,\ n-1},\quad m=0,\ 1,$$

Then

$$\varphi(\mathbf{x}_m+\beta_m\mathbf{e}_{i+1})-\varphi(\mathbf{x}_m)=\beta_m\langle\varphi'(\mathbf{x}_m+\theta_1\beta_m\mathbf{e}_{i+1}),\ \mathbf{e}_{i+1}\rangle\geqslant 0,$$

$$i=\overline{0,\ n-1},\quad m=0,\ 1,\ \ldots,$$

$$\varphi(\mathbf{x}_m - \beta_m \mathbf{e}_{i+1}) - \varphi(\mathbf{x}_m) = \beta_m \langle \varphi'(\mathbf{x}_m - \theta_2 \beta_m \mathbf{e}_{i+1}), \ -\mathbf{e}_{i+1} \rangle \geqslant 0,$$
$$i = \overline{0, \ n-1}, \quad m = 0, \ 1, \ \ldots .$$

The sequence $\{\mathbf{x}_m\}$ (remember that it belongs to the bounded set X_0) contains a convergent subsequence $\{\mathbf{x}_{m_j}\}$, $\lim\limits_{j \to \infty} \mathbf{x}_{m_j} = \bar{\mathbf{x}}$. Since $\lim\limits_{j \to \infty} \beta_{m_j} = 0$ and the function $\varphi'(\mathbf{x})$ is continuous, passing to the limit in the above inequalities we derive

$$\langle \varphi'(\bar{\mathbf{x}}), \ \mathbf{e}_{i+1} \rangle \geqslant 0, \quad \langle \varphi'(\bar{\mathbf{x}}), \ -\mathbf{e}_{i+1} \rangle \geqslant 0, \quad i = \overline{0, \ n-1}.$$

Thus, $\varphi'(\mathbf{x}) = 0$ at any limiting point of the sequence $\{\mathbf{x}_k\}$, whence

$$\lim_{k \to \infty} \rho(x_k, \ X_0^*) = 0. \ \triangle$$

9.7.6. The classical scheme. In this section we present a scheme for the coordinate-wise descent method which does not possess the advantage of independence of the choice of the descent direction on the gradient of the function but provides estimates guaranteeing the rate of convergence of the minimization process of the order of $O(1/m)$ for convex functions. These estimates are of interest due to their strong dependence on the dimension n of the space E_n.

Let $\mathbf{x}_k$ be already determined; denote

$$\max_{i=1, \ 2, \ \ldots, \ n} \left| \frac{\partial \varphi(\mathbf{x}_k)}{\partial x_i} \right| = \left| \frac{\partial \varphi(\mathbf{x}_k)}{\partial x_j} \right|.$$

It can be assumed that

$$\left| \frac{\partial \varphi(\mathbf{x}_k)}{\partial x_j} \right| > 0$$

because, if otherwise, we have $\varphi'(\mathbf{x}_k) = \mathbf{0}$, and the minimization process is ended.

In the formula $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k$ the direction of descent $-\mathbf{s}_k$ is one of the directions $\mathbf{e}_j$ or $-\mathbf{e}_j$ along which the function $\varphi(\mathbf{x})$ locally decreases.

The parameter β_k is assumed to satisfy either the condition (9.12) for $0 < \lambda \leqslant \lambda_k \leqslant 1$, $k = 0, \ 1, \ \ldots$, or the condition (9.15) for $0 < \varepsilon \leqslant q_k \leqslant 1 - \varepsilon$, $k = 0, \ 1, \ \ldots$,

We estimate α_k^2:

$$\alpha_k^2 = \frac{\langle \varphi'(x_k), s_k \rangle^2}{\| \varphi'(x_k) \|^2 \| s_k \|^2} = \frac{\left(\frac{\partial \varphi(x_k)}{\partial x_j} \right)^2}{\| \varphi'(x_k) \|^2}$$

$$\geqslant \frac{\left(\frac{\partial \varphi(x_k)}{\partial x_j} \right)^2}{n \max\limits_{i=1, 2, \ldots, n} \left(\frac{\partial \varphi(x_k)}{\partial x_i} \right)^2} = \frac{1}{n}.$$

9.7.7. Convergence. If the conditions (1)-(4) in Sec. 9.4.14 hold (the conditions (5)-(7) in the same section hold by virtue of the above) and the condition 9.4.12 is fulfilled, then Theorem 9.4.15 implies that

$$\lim_{k \to \infty} \rho(x_k, X_0^*) = 0.$$

For a convex function $\varphi(x)$ of class $C^{1,1}(E_n)$, on condition that $\operatorname{diam} X_0 = \eta < \infty$, Theorem 9.4.19 (see (9.18) and (9.19)) implies the following estimates:

If (9.12) holds, then

$$\varphi(x_m) - \varphi(x^*) \leqslant \mu_0 \left(1 + C\mu_0 \frac{1}{n} \sum_{k=0}^{m-1} \lambda_k \right)^{-1}$$

$$\leqslant \mu_0 (1 + C\mu_0 \gamma m)^{-1} < \frac{1}{C\lambda} \frac{n}{m}, \quad m = 1, 2, \ldots,$$

and in case (9.15) is fulfilled,

$$\varphi(x_m) - \varphi(x^*) \leqslant \mu_0 \left(1 + C\mu_0 \frac{1}{n} \sum_{k=0}^{m-1} \gamma_k \right)^{-1}$$

$$\leqslant \mu_0 \left(1 + C\mu_0 \frac{\gamma}{n} m \right)^{-1} < \frac{1}{C\gamma} \frac{n}{m},$$

$$m = 1, 2, \ldots, \quad 0 < C \leqslant \frac{2}{L\eta^2}.$$

The Remark 9.4.20 weakening the condition of boundedness of the set X_0 also remains valid.

It should be noted that the constant factor in $1/m$ involved in the estimates for the rate of convergence of this method is n times as great as the corresponding factor in the estimates for the gradient descent method.

When the function $\varphi(\mathbf{x})$ is strongly convex, 9.4.21 implies estimates having the form

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \mu_0 \exp\left\{-C\frac{m}{n}\right\},$$

$$\| \mathbf{x}_m - \mathbf{x}^* \|^2 \leqslant \frac{2}{\rho}\mu_0 \exp\left\{-C\frac{m}{n}\right\},$$

where the constant C is selected in an appropriate manner depending on the conditions (9.12) and (9.15).

9.7.8. The method of cyclic coordinate-wise descent. In this method the direction of descent is selected in the same way as in the method of coordinate-wise descent with duplication of steplength:

$$\mathbf{s}_k = \pm \mathbf{e}_{j(k)},\ j(k) = k \pmod n + 1,\ k = 0, 1, \ldots$$

The choice of the parameter β_k is carried out either by one-dimensional minimization, and in this case it is assumed that the inequality (9.12) holds, or by the method presented in Sec. 9.4.6, and then the validity of (9.15) is assumed.

9.7.9. Assumptions *. The convergence of the method of cyclic coordinate-wise descent will be proved assuming that the function $\varphi(\mathbf{x})$ satisfies the following requirements:

(1) $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$,

(2) $X_0^* \neq \varnothing$ (see (9.16)),

(3) $\varphi(\mathbf{x})$ is bounded from below on the set X_0.

The sequence $\{\mathbf{x}_k\}$ will be assumed to satisfy the following conditions:

(4) x_0 is an arbitrary point,

(5) $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \mathbf{s}_k$, $\mathbf{s}_k = \pm \mathbf{e}_{j(k)}$, $j(k) = k \pmod n + 1$ $(k = 0, 1, \ldots)$,

(6) either the condition (9.12) holds for $1 \geqslant \lambda_k \geqslant \lambda > 0$ or the condition (9.15) is fulfilled for $1/2 \geqslant q_k \geqslant q > 0$,

(7) $\lim\limits_{k\to\infty} \| \mathbf{x}_{k+1} - \mathbf{x}_k \| = 0.$

Theorem. *If the assumptions* 9.7.9 *and the condition* 9.4.12 *hold, then*

$$\lim_{k\to\infty} \rho(\mathbf{x}_k, X_0^*) = 0. \tag{9.17}$$

* Cf. the assumptions 9.4.14.

Proof. Using Lemma 9.2.3 and the definition of β_k we obtain (by analogy with Sec. 9.7.6)

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant C\left(\frac{\partial \varphi(\mathbf{x}_k)}{\partial x_j}\right)^2 = C\langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_j\rangle^2.$$

By virtue of the monotonicity and boundedness of the sequence $\{\varphi(\mathbf{x}_k)\}$, we have

$$\lim_{k\to\infty} \langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_{j(k)}\rangle = 0.$$

Given $\varepsilon > 0$, we take an index k_0 such that

$$|\langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_{j(k)}\rangle| \leqslant L\varepsilon \text{ and } \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leqslant \frac{\varepsilon}{n} \quad \forall k \geqslant k_0.$$

Then

$$\|\mathbf{x}_{k+i} - \mathbf{x}_k\| \leqslant \|\mathbf{x}_{k+i} - \mathbf{x}_{k+i-1}\| + \ldots + \|\mathbf{x}_{k+1} - \mathbf{x}_k\| \leqslant \frac{\varepsilon}{n} i \leqslant \varepsilon \text{ for } i \leqslant n.$$

The condition $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$ implies

$$\|\varphi'(\mathbf{x}_{k+i}) - \varphi'(\mathbf{x}_k)\| \leqslant L\varepsilon, \quad i = 1,\ 2,\ \ldots,\ n.$$

Further,

$$|\langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_{k+i}\rangle| = |\langle \varphi'(\mathbf{x}_k) - \varphi'(\mathbf{x}_{k+i}),\ \mathbf{e}_{k+i}\rangle + \langle \varphi'(\mathbf{x}_{k+i}),\ \mathbf{e}_{k+i}\rangle| \leqslant \|\varphi'(\mathbf{x}_k) - \varphi'(\mathbf{x}_{k+i})\| + |\langle \varphi'(\mathbf{x}_{k+i}),\ \mathbf{e}_{k+i}\rangle| \leqslant 2L\varepsilon, \quad i = 1,\ 2,\ \ldots,\ n,$$

and, consequently,

$$\max_{i=1,\ 2,\ \ldots,\ n} |\langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_{k+i}\rangle| \leqslant 2L\varepsilon.$$

Therefore

$$\|\varphi'(\mathbf{x}_k)\| \leqslant \sqrt{n} \max_{i=1,\ 2,\ \ldots,\ n} |\langle \varphi'(\mathbf{x}_k),\ \mathbf{e}_{k+i}\rangle| \leqslant 2\sqrt{n}\, L\varepsilon,$$

whence it follows that $\lim_{k\to\infty} \varphi'(\mathbf{x}_k) = 0$.

The validity of (9.17) is proved based on Theorem 9.2.5 for $X \equiv X_0$ and $Y \equiv X_0^*$. $\triangle$

Discussion. Here it is advisable to recall Theorem 9.2.13 on conditions guaranteeing the fulfilment of the assumption (7), namely the condition that the function $\varphi(\mathbf{x})$ is linearly non-constant on the set X_0 which is bounded

(its closedness is obvious) and the condition that the sequence $\{\mathbf{x}_k\}$ is strongly lowering. We note that under these conditions the foregoing theorem remains true, however, in this case, without the assumptions (2), (3), (7) and 9.4.12. Finally, we indicate that if the set X_0^* consists of a finite number of stationary points, then Theorem 9.2.9 implies the convergence of the sequence $\{\mathbf{x}_k\}$ to a point belonging to the set X_0^*.

Remark. It is proved in Supplement 2 that the rate of convergence of the method of cyclic coordinate-wise descent with the determination of steplength from the condition (9.15) is linear relative to the functional:

$$\varphi(\mathbf{x}_m) - \varphi(\mathbf{x}^*) \leqslant \frac{C}{m},$$

where $C > 0$ depends on $\mathbf{x}_0$ and the dimension n of the space E_n.

Chapter 10

RELAXATION METHODS OF SOLVING EXTREMAL PROBLEMS WITH CONSTRAINTS

10.1. Description of the Methods

10.1.1. In this chapter we consider descent methods for the solution of the problem of minimizing a smooth function $\varphi(\mathbf{x})$ on a set X.

The first two methods, i.e. the methods of gradient projection and conditional gradient are applicable only in minimization problems on convex sets, and in case the objective function is non-convex, it can only be guaranteed that the successive approximations converge to the set of stationary points, i.e. those points where the necessary conditions for local minimum are fulfilled. When the original problem is a convex programming problem, both methods allow a priori estimates for the rate of convergence, i.e. possess definite quality characteristics.

In unconditional minimization methods no major difficulties were encountered in the determination of the direction of descent. The most time consuming part was the calculation of the gradient of the objective function. In problems with constraints, when selecting the direction of descent, it is necessary to take into account the following two factors: the direction must be feasible and must guarantee the decrease of the minimized function. In this case the choice of the direction of descent is related to the computation of $\varphi'(\mathbf{x})$ and, in a number of methods, $f_i'(\mathbf{x})$ as well, and to the solution of an extremal problem at each step. Owing to the latter, the conditional gradient method is used only for sets determined by linear constraints since in these cases to select the direction of descent it suffices to solve a linear programming problem while the gradient projection method is used for a set X of such a form that the problem of finding the projection of a point on the set is sufficient-

ly simple from the viewpoint of the numerical realization since it is the solution of this problem that determines the direction of descent.

The third method, the method of feasible directions, is characterized by the property that the choice of the direction of descent is performed by solving either a linear programming problem or a simple quadratic programming problem, generally, with a small number of constraints, and from this standpoint the method can be given preference. However, the absence of a priori estimates for the rate of convergence in convex programming problems makes it difficult to compare the method with the above-mentioned methods.

10.1.2. Conditions for extremality. We present here optimality conditions in a form suitable for the consideration of the conditional gradient and gradient projection methods.

We shall assume that the set X is convex and closed. It follows from Section 3.4.7 that the condition

$$\langle \varphi'(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^* \rangle \geqslant 0 \qquad \forall \mathbf{x} \in X \tag{10.1}$$

is necessary for $\mathbf{x}^*$ to be a point of local minimum of the function $\varphi(\mathbf{x})$ on the set X.

If the function $\varphi(\mathbf{x})$ *is convex, then the condition* (10.1) *is necessary and sufficient for* $\mathbf{x}^* = \arg\min\{\varphi(\mathbf{x}): \mathbf{x} \in X\}$.

Indeed, the necessity follows from the above and the sufficiency is implied by the condition of convexity: $\varphi(\mathbf{x}) - \varphi(\mathbf{x}^*) \geqslant \langle \varphi'(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^* \rangle \geqslant 0$. $\triangle$

Further, let $\mathbf{p}^*$ be the projection of the point $\mathbf{v}^* = \mathbf{x}^* - \nu\varphi'(\mathbf{x}^*)$, $\nu > 0$, on the set X. The condition

$$\mathbf{x}^* = \mathbf{p}^* \tag{10.2}$$

is necessary for $\mathbf{x}^*$ to be a point of local minimum of the function $\varphi(\mathbf{x})$ on the set X. Indeed, (10.1) implies $-\nu\langle \varphi'(\mathbf{x}^*),\ \mathbf{x} - \mathbf{x}^* \rangle \leqslant 0$. However, $-\nu\varphi'(\mathbf{x}^*) = \mathbf{v}^* - \mathbf{x}^*$, and, consequently, $\langle \mathbf{v}^* - \mathbf{x}^*,\ \mathbf{x} - \mathbf{x}^* \rangle \leqslant 0$, i.e. by virtue of Theorem 2.2.3, $\mathbf{x}^*$ is the projection of the point $\mathbf{v}^*$ on X^*: $\mathbf{x}^* = \mathbf{p}^*$. $\triangle$

If the function $\varphi(\mathbf{x})$ *is convex, then* (10.2) *is a necessary and sufficient condition for*

$$\mathbf{x}^* = \arg\min\{\varphi(\mathbf{x}): \quad \mathbf{x} \in X\}.$$

The *necessity* follows from the above.

Sufficiency. Let $\mathbf{x}^* = \mathbf{p}^*$. From Theorem 2.2.3 we derive

$$\langle \mathbf{x} - \mathbf{x}^*, \mathbf{v}^* - \mathbf{x}^* \rangle \leqslant 0 \quad \forall \mathbf{x} \in X.$$

Since $\mathbf{v}^* = \mathbf{x}^* - \nu\varphi'(\mathbf{x}^*)$ and $\nu > 0$, we have $\langle \varphi'(\mathbf{x}^*), \mathbf{x} - \mathbf{x}^* \rangle \geqslant 0$, i.e. $\mathbf{x}^*$ is a point of minimum. $\triangle$

10.2. Gradient Projection Method

10.2.1. The idea of the method. As was already mentioned, the gradient projection method is a widely applied relaxation method for solving unconditional minimization problems. However, in the case of problems with constraints the direction along the antigradient is not necessarily feasible. If the set X is convex, then, when finding the direction of descent at the point $\mathbf{x}_k$, it appears natural to project the point $\mathbf{v}_k = \mathbf{x}_k - \nu_k\varphi'(\mathbf{x}_k)$ (where ν_k is a fixed positive number) on the set X and take, as the direction of descent, the vector $-\mathbf{s}_k = \mathbf{p}_k - \mathbf{x}_k$ where $\mathbf{p}_k$ is the projection of $\mathbf{v}_k$ on X, after which the descent along this direction is carried out. In this case the convexity of the set X guarantees that $\mathbf{p}_k - \mathbf{x}_k$ is a feasible direction.

Hence, the gradient projection method consists in calculating the projection $\mathbf{p}_k$ of the point $\mathbf{v}_k = \mathbf{x}_k - \nu_k\varphi'(\mathbf{x}_k)$ on the set X and selecting the parameter β_k in such a way that $\varphi(\mathbf{x}_{k+1}) < \varphi(\mathbf{x}_k)$ at the point $\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k(\mathbf{x}_k - \mathbf{p}_k)$.

To determine the direction $\mathbf{s}_k = \mathbf{x}_k - \mathbf{p}_k$ one should solve the problem of minimizing the quadratic function $\| \mathbf{v}_k - \mathbf{x} \|^2$ on the set X. In the general case the complexity of this problem is of the same order as in the original problem. However, for problems where the feasible region has a simple geometrical structure, e.g. is a multidimensional parallelepiped, the determination of $\mathbf{p}_k$ reduces to comparing n numbers.

In the gradient projection method the point $\mathbf{p}_k$ can be calculated approximately, i.e. instead of $\mathbf{p}_k$ a point $\mathbf{y}_k$ is found and $\mathbf{x}_{k+1}$ is constructed by the formula

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k(\mathbf{x}_k - \mathbf{y}_k), \tag{10.3}$$

however, here the following conditions must be fulfilled:

$$\mathbf{y}_k \in X, \tag{10.4}$$

$$\varphi(\mathbf{x}_k - \beta_k(\mathbf{x}_k - \mathbf{y}_k)) \leqslant (1 - \lambda_k)\varphi(\mathbf{x}_k) + \lambda_k \omega_k, \tag{10.5}$$

$$0 < \lambda \leqslant \lambda_k \leqslant 1, \quad 0 \leqslant \beta_k \leqslant 1, \tag{10.6}$$

where

$$\omega_k = \min_{\beta \in [0,\,1]} \varphi(\mathbf{x}_k - \beta(\mathbf{x}_k - \mathbf{p}_k)). \tag{10.7}$$

We stress that the choice of β_k and $\mathbf{y}_k$ in the construction of the sequence $\{\mathbf{x}_k\}$ is rather arbitrary and it is only required that the conditions (10.4)-(10.6) be satisfied.

10.2.2. The set of stationary points. Consider the set

$$X^* = \{\mathbf{x}^* \in X: \| \mathbf{x}^* - \mathbf{p}^* \| = 0\}$$

of stationary points of a function $\varphi(\mathbf{x})$ on a set X, i.e. the points where the necessary conditions for local minimum are fulfilled. We also introduce the sets $X_0 = \{\mathbf{x} \in X: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$ and $X_0^* = X^* \cap X_0$.

10.2.3. The condition guaranteeing convergence. Given a number $\varepsilon > 0$, we consider the set $U_\varepsilon = \{\mathbf{x} \in X: \rho(\mathbf{x}, X_0^*) \leqslant \varepsilon\}$, where

$$\rho(\mathbf{x}, X_0^*) = \inf_{\mathbf{x}^* \in X_0^*} \| \mathbf{x}^* - \mathbf{x} \|^2.$$

The condition guaranteeing the convergence of the relaxation sequence consists in the following: *for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that for all* $\mathbf{x} \in X_0 \setminus U_\varepsilon$ *there holds* $\| \mathbf{x} - \mathbf{p} \| \geqslant \delta$ *for* $\nu > 0$.

10.2.4. Assumptions. The function $\varphi(\mathbf{x})$ is assumed to satisfy the following conditions:

(1) $\varphi(\mathbf{x}) \in C^{1,1}(X)$,

(2) $X_0^* \neq \varnothing$,

(3) $\varphi(\mathbf{x})$ is bounded from below on the set X_0.

The sequence $\{\mathbf{x}_k\}$ is assumed to possess the following properties:

(4) $\mathbf{x}_0$ is an arbitrary point belonging to X,

(5) the conditions (10.3)-(10.6) hold for $0 < \nu' \leqslant \nu_k \leqslant \nu'' < \infty$.

10.2.5. Theorem. *If the assumptions* (10.2.4) *and the condition* 10.2.3 *hold, then* $\lim_{k \to \infty} \rho(\mathbf{x}_k, X_0^*) = 0$.

Proof. From Lemma 9.2.3 and the condition (10.5) it follows that for any $\beta \in [0, 1]$ we have

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k(\varphi(x_k) - \omega_k) \geqslant \lambda_k(\varphi(x_k) - \varphi(x_k - \beta(x_k - p_k))) \geqslant \lambda_k(\beta \langle \varphi'(x_k), x_k - p_k \rangle - L\beta^2 \| x_k - p_k \|^2/2).$$

Section 2.2.4 implies that for any $\nu_k > 0$ the inequality

$$\nu_k \langle \varphi'(x_k), x_k - p_k \rangle \geqslant \| x_k - p_k \|^2 \qquad (10.8)$$

takes place. Therefore

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k(\beta/\nu_k - L\beta^2/2) \| x_k - p_k \|^2. \qquad (10.9)$$

Taking $\beta = \min\{1, 1/\nu_k L\}$ we obtain for $\beta = 1$ (in view of the fact that in this case $L \leqslant 1/\nu_k$) the relation

$$\frac{1}{\nu_k}\beta - \frac{1}{2}L\beta^2 = \frac{1}{\nu_k} - \frac{1}{2}L \geqslant \frac{1}{2\nu_k} \geqslant \frac{1}{2\nu''}.$$

In case $\beta = 1/\nu_k L \leqslant 1$, we have

$$\frac{1}{\nu_k}\beta - \frac{1}{2}L\beta^2 = \frac{1}{2\nu_k^2 L} \geqslant \frac{1}{2(\nu'')^2 L}.$$

Finally, we arrive at the inequality

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda\Delta \| x_k - p_k \|^2, \qquad (10.10)$$

where $\Delta = \frac{1}{2\nu''} \min\{1, 1/\nu'' L\}$.

The sequence $\{\varphi(x_k)\}$ is monotonic and bounded from below, and therefore $\varphi(x_k) - \varphi(x_{k+1}) \to 0$ for $k \to \infty$ and $f(x_k) \equiv \| x_k - p_k \|^2 \to 0$ for $k \to \infty$.

The proof of the theorem follows from Theorem 9.2.5 for $X \equiv X_0$ and $Y = X_0^*$. $\triangle$

10.2.6. Theorem. *If*

(a) $\varphi(x)$ *is a convex and differentiable function on a convex and closed set* X,

(b) $\operatorname{diam} X_0 = \eta < \infty$,

(c) $\{x_k\}$ *is a relaxation sequence,*

then the estimate

$$\varphi(x_m) - \varphi^* \leqslant \mu_0 \left[1 + C\mu_0 \sum_{k=0}^{m-1} \frac{\varphi(x_k) - \varphi(x_{k+1})}{\| x_k - p_k \|^2} \right]^{-1}, \quad m = 1, 2, \ldots \qquad (10.11)$$

takes place.

Proof. We first perform the following identical transformations:

$$\langle \varphi'(x_k),\ x_k - x^* \rangle = \langle \varphi'(x_k),\ x_k - p_k \rangle + \langle \varphi'(x_k),\ p_k - x^* \rangle$$
$$= \langle \varphi'(x_k),\ x_k - p_k \rangle + \frac{1}{\nu_k} \langle x_k - p_k,\ x^* - p_k \rangle$$
$$- \frac{1}{\nu_k} \langle x_k - p_k,\ x^* - p_k \rangle.$$

This and the condition that $\langle x_k - p_k,\ x^* - p_k \rangle \leqslant 0$ imply

$$\langle \varphi'(x_k),\ x_k - x^* \rangle$$
$$\leqslant \langle \varphi'(x_k),\ x_k - p_k \rangle - \frac{1}{\nu_k} \langle x_k - p_k,\ x^* - p_k \rangle$$
$$\leqslant \left(\| \varphi'(x_k) \| + \frac{1}{\nu_k} \| x^* - p_k \| \right) \| x_k - p_k \|$$
$$\leqslant \left(\| \varphi'(x_k) \| + \frac{1}{\nu_k} \| x_k - x^* \| \right.$$
$$\left. + \frac{1}{\nu_k} \| x_k - p_k \| \right) \| x_k - p_k \|.$$

However, (10.8) implies the inequality $\|x_k - p_k\|/\nu_k \leqslant \|\varphi'(x_k)\|$. In view of $\|x^* - x_k\| \leqslant \eta$ and $\|\varphi'(x_k)\| \leqslant \gamma < \infty$, we derive the final inequalities

$$0 \leqslant \langle \varphi'(x_k),\ x_k - x^* \rangle \leqslant (2\gamma + \eta/\nu_k) \| x_k - p_k \|$$

holding for any point $x^* \in X^*$ and, in particular, for $x^* = p^*(x_k)$, the projection of the point x_k on the set X^*. This and the inequality (see 9.3.2)

$$\varphi(x_m) - \varphi^* \leqslant \mu_0 \left[1 + \mu_0 \sum_{k=0}^{m-1} \frac{\varphi(x_k) - \varphi(x_{k+1})}{\langle \varphi'(x_k),\ x_k - p^*(x_k) \rangle^2} \right]^{-1}$$

imply (10.11) for $0 < C \leqslant (2\gamma + \eta/\nu')^{-2}$. $\triangle$

10.2.7. Estimation of the rate of convergence in the convex case.

Theorem. *If the set X is convex and closed, and*

(a) *the function $\varphi(x)$ on X is convex and belongs to the class $C^{1,1}(X)$,*

(b) $\operatorname{diam} X_0 = \eta < \infty$,

(c) *the sequence $\{x_k\}$ is determined by the conditions (10.3)–(10.6) for $0 \leqslant \nu' \leqslant \nu_k \leqslant \nu''$,*

then the estimate

$$\varphi(x_m) - \varphi^* \leqslant \mu_0 [1 + \mu_0 C m]^{-1}, \quad m = 1, 2, \ldots, \tag{10.12}$$

holds for any

$$0 < C \leqslant \frac{\lambda}{2\nu''} \left(2\gamma + \frac{\eta}{\nu'}\right)^{-2} \min\left\{1, \frac{1}{\nu'' L}\right\},$$

where $\gamma = \sup\limits_{x \in X_0} \| \varphi'(x) \|$.

The proof follows from the inequalities (10.10) and (10.11). $\triangle$

10.2.8. On the ways of selecting the steplength. Most of what was said about the choice of the steplength in Sec. 9.4.3, 9.4.4, and 9.4.5 remains valid, however with regard to $\beta_k \in [0, 1]$. Thus, to calculate the parameter β_k one usually applies one of the one-dimensional minimization methods. The procedure of selecting β_k can be substantially simplified when the Lipschitz constant L in the condition $\| \varphi'(x) - \varphi'(y) \| \leqslant L \| x - y \|$ is known. In this case one can take as β_k any of the numbers satisfying the condition $0 < \bar{\beta} \leqslant \beta_k \leqslant 1$; however, here the parameter ν_k must satisfy the inequalities $0 \leqslant \nu' \leqslant \nu_k \leqslant \nu'' < 2/L$. Then (10.9) implies

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \bar{\beta} \left(\frac{1}{\nu'} - \frac{1}{2} L\right) \| x_k - p_k \|^2,$$

and the convergence and the estimate for the rate of convergence are retained by virtue of Sec. 10.2.5 and 10.2.6, respectively. There are also some other ways of determining the parameter β_k based on the knowledge of the Lipschitz constant L. For instance, $\beta_k = \min\{1, \rho_k/\nu_k\}$ where $0 < \varepsilon_1 \leqslant \rho_k \leqslant (2 - \varepsilon_2)/L$ and $0 < \nu' \leqslant \nu_k \leqslant \nu'' < \infty$. It can easily be seen that both the convergence and the estimate for its rate are retained in this case as well.

10.2.9. Conditions for selecting the parameter β_k implementable without one-dimensional minimization procedure. As β_k we shall take $\beta_k = \beta_k^*$, i.e. the greatest of the numbers satisfying the inequalities

$$\varphi(x_k) - \varphi(x_k - \beta_k (x_k - p_k)) \geqslant \frac{1}{2\nu_k} \beta_k \| x_k - p_k \|^2, \quad 0 \leqslant \beta_k \leqslant 1. \tag{10.13}$$

We shall show that there is $\beta_k \in [0, 1]$ satisfying (10.13). Indeed, if one takes $\bar{\beta}_k = \min\{1, 1/\nu_k L\}$, then (see the derivation of (10.10))

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \bar{\beta}_k(\mathbf{x}_k - \mathbf{p}_k)) \geqslant \bar{\beta}_k\left(\frac{1}{\nu_k} - \frac{1}{2}L\bar{\beta}_k\right)\|\mathbf{x}_k - \mathbf{p}_k\|^2$$

$$\geqslant \frac{1}{2\nu_k}\min\left\{1, \frac{1}{\nu_k L}\right\}\|\mathbf{x}_k - \mathbf{p}_k\|^2 = \frac{1}{2\nu_k}\bar{\beta}_k\|\mathbf{x}_k - \mathbf{p}_k\|^2.$$

Since the problem of finding the maximum number satisfying the inequality (10.13) is solved approximately, it is possible to write down relations for the parameter β_k determined as the maximum value for some $q_k \in (0, 1/2]$:

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta_k(\mathbf{x}_k - \mathbf{p}_k)) \geqslant \frac{q_k}{\nu_k}\beta_k\|\mathbf{x}_k - \mathbf{p}_k\|^2, \quad 0 \leqslant \beta_k \leqslant 1. \tag{10.14}$$

It is obvious that both the convergence and the estimate for its rate are retained in this case as well.

10.2.10. On checking the convergence. To check the convergence during the calculation process carried out according to the gradient projection method it is advisable to compute at each iteration the ratio $(\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}))/\|\mathbf{x}_k - \mathbf{p}_k\|^2$. The estimate (10.11) indicates that in the convex case the process converges at a rate of the order not lower than $O(1/m)$ as long as $(\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}))/\|\mathbf{x}_k - \mathbf{p}_k\|^2 \geqslant \varepsilon$, where the parameter ε is determined by the accepted calculation accuracy.

10.3. Method of Conditional Gradient

10.3.1. The idea of the method. The choice of the direction of descent in this method consists in the following: the function $\varphi(\mathbf{x})$ is linearized at the point $\mathbf{x}_k$ by constructing the linear expression $\varphi_L(\mathbf{x}) = \varphi(\mathbf{x}_k) + \langle\varphi'(\mathbf{x}_k), \mathbf{x} - \mathbf{x}_k\rangle$ and then $\varphi_L(\mathbf{x})$ is minimized on the set X and the point $\mathbf{y}_k$ is found. After that one puts $-\mathbf{s}_k = \mathbf{y}_k - \mathbf{x}_k$ and then descent along the direction $-\mathbf{s}_k$ is performed. Hence, to determine the direction $-\mathbf{s}_k$ the problem of minimizing a linear function on the set X should be solved. Generally, this is a problem of the same degree of

complexity as the original one. However, when the feasible region is determined by linear constraints, it becomes a linear programming problem finite methods of whose solution were considered in Chap. 5.

We remind the reader that the set X is assumed to be convex and closed and the function $\varphi(\mathbf{x})$ is assumed to be of class $C^{1,1}(X)$.

10.3.2. The scheme of the method. An arbitrary point belonging to the set X is taken as $\mathbf{x}_0$. The point $\mathbf{x}_{k+1}$ ($k = 1, 2, \ldots$) is constructed according to the formula

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k (\mathbf{x}_k - \mathbf{y}_k). \tag{10.3}$$

Here $\mathbf{y}_k$ is chosen so that

$$\langle \varphi'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{y}_k \rangle \geqslant \min \{\sigma_k \sup_{\mathbf{x} \in X} \langle \varphi'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x} \rangle, \alpha_k\}, \tag{10.15}$$

where α_k are given numbers such that $\alpha_k \geqslant \alpha > 0$ and the parameters σ_k satisfy the inequality $0 < \sigma \leqslant \sigma_k \leqslant 1$. Moreover,

$$\| \mathbf{x}_k - \mathbf{y}_k \| \leqslant \varkappa < \infty. \tag{10.16}$$

The parameters β_k are found by minimizing the function $\psi(\beta) = \varphi(\mathbf{x}_k - \beta(\mathbf{x}_k - \mathbf{y}_k))$ on the closed interval $[0, 1]$.

We shall assume that β_k satisfies

$$\varphi(\mathbf{x}_k - \beta_k (\mathbf{x}_k - \mathbf{y}_k)) \leqslant (1 - \lambda_k) \varphi(\mathbf{x}_k) + \lambda_k \omega_k, \tag{10.5}$$

$$0 < \lambda \leqslant \lambda_k \leqslant 1, \quad 0 \leqslant \beta_k \leqslant 1, \tag{10.6}$$

where

$$\omega_k = \min_{\beta \in [0, 1]} \varphi(\mathbf{x}_k - \beta(\mathbf{x}_k - \mathbf{y}_k)). \tag{10.17}$$

10.3.3. Discussion. The condition that $\mathbf{y}_k$ is found from (10.15) can be illustrated by the following example. If the set X is determined by linear constraints, the problem of finding $\mathbf{y}_k$ becomes a linear programming problem of maximizing the linear function $\langle \varphi'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x} \rangle$ on the polyhedral set X. In the solution of this problem the conditions (10.15) make it possible to confine oneself to several simplex steps in the direction of increase of the function $\langle \varphi'(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x} \rangle$.

Further, the set X may turn out to be unbounded. In this case, despite the fact that the original problem is solvable, the problem of selecting the direction of descent can have no solution, i.e. there can occur a situation where $\sup_{x \in X} \langle \varphi'(x_k), x_k - x \rangle = +\infty$ for some index k. The condition (10.15) allows one to find the point y_k in this case as well. It should be noted that the choice of the numbers α_k and the appearance of the parameters σ_k are limited by the condition (10.16). The condition guaranteeing convergence remains the same as in the gradient projection method (see Sec. 10.2.3).

10.3.4. Assumptions. The function $\varphi(x)$ is assumed to have the following properties:

(1) $\varphi(x) \in C^{1,1}(X)$,

(2) $X_0^* \neq \varnothing$,

(3) $\varphi(x)$ is bounded from below on the set X_0.

The sequence $\{x_k\}$ is assumed to satisfy the following conditions:

(4) x_0 is an arbitrary point belonging to X,

(5) the relation (10.3), (10.5), (10.6), (10.15), and (10.16) hold.

10.3.5. Theorem. *If the assumptions* 10.3.4 *and the condition* 10.2.3 *hold, then*

$$\lim_{k \to \infty} \rho(x_k, X_0^*) = 0.$$

Proof. Lemma 9.2.3 and the condition (10.5) imply, for any $\beta \in [0, 1]$, the inequalities

$$\begin{aligned}\varphi(x_k) - \varphi(x_{k+1}) &\geqslant \lambda_k (\varphi(x_k) - \omega_k) \geqslant \lambda_k (\varphi(x_k) - \varphi(x_k - \beta(x_k - y_k))) \\ &\geqslant \lambda_k (\beta \langle \varphi'(x_k), x_k - y_k \rangle - L\beta^2 \| x_k - y_k \|^2/2).\end{aligned}$$

Taking

$$\beta = \min \left\{ 1, \frac{\langle \varphi'(x_k), x_k - y_k \rangle}{L \| x_k - y_k \|^2} \right\}$$

we see that

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{\lambda \langle \varphi'(x_k), x_k - y_k \rangle}{2} \qquad (10.18)$$

for $\beta = 1 \leqslant \dfrac{\langle \varphi'(x_k), x_k - y_k \rangle}{L \| x_k - y_k \|^2}$

and

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant \frac{1}{2L} \frac{\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{y}_k\rangle^2}{\|\mathbf{x}_k - \mathbf{y}_k\|^2} \geqslant \frac{\lambda}{2L\varkappa^2} \langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{y}_k\rangle^2 \qquad (10.19)$$

for $\beta = \dfrac{\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{y}_k\rangle}{L\|\mathbf{x}_k - \mathbf{y}_k\|^2}$.

Let $\mathbf{p}_k$ denote the projection of the point $\mathbf{v}_k = \mathbf{x}_k - \varphi'(\mathbf{x}_k)$ on the set X. Then, by virtue of (10.8), $\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k\rangle \geqslant \|\mathbf{x}_k - \mathbf{p}_k\|^2$, and taking into consideration that $\sigma_k \geqslant \sigma > 0$ and $\alpha_k \geqslant \alpha > 0$, we find (see (10.15))

$$\begin{aligned}\langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{y}_k\rangle &\geqslant \min\{\sigma_k \sup_{\mathbf{x}\in X} \langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{x}\rangle,\ \alpha_k\} \\ &\geqslant \min\{\sigma_k \langle \varphi'(\mathbf{x}_k),\ \mathbf{x}_k - \mathbf{p}_k\rangle,\ \alpha_k\} \\ &\geqslant \min\{\sigma_k \|\mathbf{x}_k - \mathbf{p}_k\|^2,\ \alpha_k\} \geqslant \min\{\sigma \|\mathbf{x}_k - \mathbf{p}_k\|^2,\ \alpha\}.\end{aligned} \qquad (10.20)$$

Since the sequence $\{\varphi(\mathbf{x}_k)\}$ is monotonic and bounded from below, we have $\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \to 0$ for $k \to \infty$, and, consequently, (10.18), (10.19) and (10.20) imply that $\|\mathbf{x}_k - \mathbf{p}_k\| \to 0$ for $k \to \infty$. Consequently, there is an index k_0 such that $\sigma\|\mathbf{x}_k - \mathbf{p}_k\|^2 < \min\{1, \alpha\}$ for all $k \geqslant k_0$, and therefore from (10.18)-(10.20) we derive the inequality

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant C\|\mathbf{x}_k - \mathbf{p}_k\|^2 \quad \forall k \geqslant k_0,$$

where $C = \dfrac{\lambda}{2}\sigma \min\{1,\ 1/L\varkappa^2\}$.

The proof of the theorem now follows from Theorem 9.2.5 for $X = X_0$ and $Y = X_0^*$. $\triangle$

10.3.6. Estimating the rate of convergence in the convex case.

Theorem. *If the set X is convex and closed, and*

(a) *the function $\varphi(\mathbf{x})$ on X is convex and belongs to the class $C^{1,1}(X)$,*

(b) $\operatorname{diam} X_0 = \eta < \infty$,

(c) *the sequence $\{\mathbf{x}_k\}$ is determined by the conditions* (10.3), (10.5), (10.6), (10.15), *and* (10.16),

then the estimate

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \mu_0[1 + \mu_0 C m]^{-1}, \quad m = 1, 2, \ldots, \qquad (10.21)$$

holds for any

$$0<C\leqslant\frac{\lambda}{2\varkappa}\min\left\{\frac{1}{\gamma},\ \frac{1}{L}\right\}\left[\min\left\{\sigma,\ \frac{\alpha}{\gamma\eta}\right\}\right]^2.$$

Proof. Since $\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle>0$ (if $\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle=0$, we have $\mathbf{x}_k=\mathbf{p}_k$, and $\mathbf{x}_k$ is a point of minimum) and $\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle\leqslant\|\varphi'(\mathbf{x}_k)\|\,\|\mathbf{x}_k-\mathbf{y}_k\|\leqslant\gamma\varkappa$, where $\gamma=\sup\limits_{\mathbf{x}\in X_0}\|\varphi'(\mathbf{x})\|$, we conclude from (10.18) that

$$\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})\geqslant\frac{\lambda}{2\gamma\varkappa}\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle^2,$$

and, by virtue of (10.19), arrive at the inequality

$$\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})\geqslant C_1\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle^2 \quad (10.22)$$

for $C_1=\dfrac{\lambda}{2\varkappa}\min\{1/\gamma,\ 1/L\varkappa\}$.

If $\sigma_k\sup\limits_{\mathbf{x}\in X}\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}\rangle\leqslant\alpha_k$, then (10.15) implies the inequality

$$\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle\geqslant\sigma_k\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle\geqslant\sigma\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle.$$

In case $\sigma_k\sup\limits_{\mathbf{x}\in X}\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}\rangle>\alpha_k$, we have

$$\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle\geqslant\alpha_k\geqslant\frac{\alpha_k}{\gamma\eta}\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle\geqslant\frac{\alpha}{\gamma\eta}\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle.$$

Let $C_2=\min\{\sigma,\ \alpha/(\gamma\eta)\}$. Then $\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{y}_k\rangle\geqslant C_2\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle$. Finally, this and (10.22) imply the inequality

$$\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})\geqslant C\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{x}^*\rangle^2,\quad C=C_1C_2^2, \quad (10.23)$$

holding for any point $\mathbf{x}^*\in X^*$ and, in particular, for $\mathbf{x}^*=\mathbf{p}^*(\mathbf{x}_k)$, the projection of the point $\mathbf{x}_k$ on the set X^*. From (10.23) and the estimate (see 9.3.2).

$$\varphi(\mathbf{x}_m)-\varphi^*\leqslant\mu_0\left[1+\mu_0\sum_{k=0}^{m-1}\frac{\varphi(\mathbf{x}_k)-\varphi(\mathbf{x}_{k+1})}{\langle\varphi'(\mathbf{x}_k),\ \mathbf{x}_k-\mathbf{p}^*(\mathbf{x}_k)\rangle^2}\right]^{-1}$$

we obtain (10.21). Δ

10.3.7. Discussion. It is clear that, with decreasing C, the a priori estimate for the rate of convergence of the method worsens. To elucidate the question of selecting α_k we estimate the constant C: under some natural assumptions the inequality $C \leqslant \bar{C}\alpha^2/\varkappa^2$ holds, where $\bar{C} = \text{const} > 0$. Since $\| x_k - y_k \|$ depends on the choice of α, and, consequently, $\varkappa$ depends on α, the relations estimating the magnitude of C show that it is advisable to choose α_k so that the ratio $\alpha_k^2 / \| x_k - y_k \|$ does not decrease with increasing index k.

10.4. Method of Feasible Directions

10.4.1. The idea of the method. In practice, to determine the points of local minima in the general nonlinear programming problem use can be made of those relaxation methods where the calculation of the direction of descent from a point x takes into account local properties of the function under minimization and local properties of the set X in the neighborhood of the point x. Among the known methods, the method of feasible directions satisfies these conditions. The direction of descent in this method is found in the following way: one constructs at the point x the cone of feasible directions which is determined by a linear system of inequalities and takes in the cone a vector of finite length forming an acute angle with the antigradient of the objective function. Thus, the method of feasible directions can be regarded as a natural extension of the gradient descent method to minimization problems with constraints. In contrast to the two foregoing methods, the method of feasible directions possesses an advantage that to determine the direction of descent it suffices to solve a linear programming problem with a small number of constraints or a simplest quadratic programming problem.

Consider the problem

$$\varphi(x) \to \min$$

on the set $X = \{x: f_i(x) \geqslant 0, \quad i = \overline{1, m}\}$.

The relaxation sequence $\{x_k\} \subset X$ is constructed by the formula

$$x_{k+1} = x_k - \beta_k s_k, \quad k = 0, 1, \ldots \tag{10.24}$$

10.4.2. Determining the direction of descent. We consider the problem of finding a number σ and a vector $\mathbf{s}$ such that

$$\begin{aligned} \sigma &\to \max, \\ \langle f_i'(\mathbf{x}),\ \mathbf{s}\rangle + \sigma &\leqslant 0, \quad i \in I, \\ -\langle \varphi'(\mathbf{x}),\ \mathbf{s}\rangle + \sigma &\leqslant 0, \\ \langle \mathbf{s},\ \mathbf{s}\rangle &\leqslant 1. \end{aligned} \tag{10.25}$$

The last constraint is nonlinear; it excludes the possibility that the solution to the problem (10.25) can be unbounded. This constraint is sometimes replaced by the following: $-1 \leqslant s_j \leqslant 1$, $j = \overline{1, n}$, and then (10.25) becomes a linear programming problem. In our further consideration we confine ourselves to the problem (10.25); however, all that was said above applies to the problem (10.25) with linear constraints as well.

Denote by $\bar{\sigma}(x, \varepsilon)$, $\bar{\mathbf{s}}(x, \varepsilon)$ a solution to the problem (10.25) for $I = I(\mathbf{x}, \varepsilon) = \{i: 0 \leqslant f_i(\mathbf{x}) \leqslant \varepsilon\}$, where ε is a positive number and $\bar{\sigma}(\mathbf{x}, 0)$ and $\bar{\mathbf{s}}(\mathbf{x}, 0)$ are the solutions to the problem (10.25) for $I = I(\mathbf{x}, 0) = \{i: f_i(\mathbf{x}) = 0\}$.

For a direction $-\mathbf{s}$, $\|\mathbf{s}\| = 1$, *at a point* $\mathbf{x} \in X$ *to be feasible it is sufficient that* $\sigma > 0$, *and* $\mathbf{s}$ *satisfy the inequalities* $\langle f_i'(\mathbf{x}),\ \mathbf{s}\rangle + \sigma \leqslant 0$ *for all* $i \in I(\mathbf{x}, \varepsilon)$ *and for at least one* $\varepsilon > 0$.

The proof of this assertion is analogous to that of Theorem 3.2.6. Hence, the first group of constraints with $i \in I$ in the problem (10.25) guarantees that the direction $-\mathbf{s}$ is feasible at the point $\mathbf{x}$ and the last constraint implies that $\varphi(\mathbf{x})$ decreases along this direction.

10.4.3. The distance to the boundary. We introduce the parameter ζ equal to the distance from a point $\mathbf{x}$ to the nearest boundary point of the set X in the direction $-\mathbf{s}$. If $-\mathbf{s}$ is a feasible direction at the point $\mathbf{x}$, then there is a number $\bar{\beta} > 0$ such that the point $\mathbf{x} - \beta\mathbf{s}$ belongs to X for all $\beta \in [0, \bar{\beta}]$. In this case $\zeta = \sup \bar{\beta}$ (if the supremum is finite) is equal to the longest line segment $[\mathbf{x}, \mathbf{x} - \zeta\mathbf{s}]$ entirely belonging to the set X. Here $\mathbf{y} = \mathbf{x} - \zeta\mathbf{s}$ is a boundary point of the set X. When $\zeta = +\infty$, the set X contains the whole ray $\mathbf{x} - \beta\mathbf{s}$, $\beta \geqslant 0$. For $\mathbf{x} = \mathbf{x}_k$ and $\mathbf{s} = \mathbf{s}_k$ we shall denote $\zeta = \zeta_k$.

10.4.4. The scheme of the method. As an initial approximation x_0, an arbitrary point belonging to the set X can be selected, and ε_0 is taken in the half-open interval $(0, 1]$. Let the kth iteration result in the calculation of x_k and ε_k. We shall describe the $(k+1)$th iteration.

Step A. The problem (10.25) is solved for $I = I(x_k, \varepsilon_k)$ to calculate feasible σ_k and s_k, $\|s_k\| = 1$, such that $\sigma_k \geqslant \xi_k \bar{\sigma}(x_k, \varepsilon_k)$, where $0 < \xi \leqslant \xi_k \leqslant 1$.

Step B. If $\sigma_k \geqslant \varepsilon_k$, one passes to the calculation of the parameter β_k. Usually β_k is calculated by solving the one-dimensional problem of minimizing the function $\psi(\beta) = \varphi(x_k - \beta s_k)$ for $0 \leqslant \beta \leqslant \zeta_k$. In this case the parameter β_k must satisfy the conditions

$$\varphi(x_k - \beta_k s_k) \leqslant (1 - \lambda_k)\varphi(x_k) + \lambda_k \omega_k,$$

$$0 < \lambda \leqslant \lambda_k \leqslant 1, \quad \omega_k = \inf_{0 \leqslant \beta \leqslant \zeta_k} \varphi(x_k - \beta s_k). \quad (10.26)$$

The parameters β_k can also be selected in the following way. Let $\bar{\beta}_k$ be the greatest of the numbers satisfying the relations

$$\varphi(x_k) - \varphi(x_k - \beta_k s_k) \geqslant \frac{1}{2}\beta_k \sigma_k, \quad 0 \leqslant \beta_k \leqslant \zeta_k. \quad (10.27)$$

As β_k one can take any number satisfying the inequalities (10.27) and the condition $\beta_k \geqslant \alpha \bar{\beta}_k$ for any $\alpha \in [0, 1]$.

Finally, one calculates $x_{k+1} = x_k - \beta_k s_k$, puts $\varepsilon_{k+1} = \varepsilon_k$, and passes to step A.

Step C. If $0 < \sigma_k < \varepsilon_k$, then one puts $x_{k+1} = x_k$, $\varepsilon_{k+1} = \gamma_k \varepsilon_k$, where $0 < \gamma_k \leqslant \gamma < 1$, and passes to step A. If $\sigma_k = 0$, one calculates $\bar{\sigma}(x_k, 0)$ by solving the problem (10.25) for $I = I(x_k, 0)$. If $\bar{\sigma}(x_k, 0) = 0$, the process is stopped. If otherwise, one puts $x_{k+1} = x_k$, $\varepsilon_{k+1} = \gamma_k \varepsilon_k$ where $0 < \gamma_k \leqslant \gamma < 1$ and passes to step A.

10.4.5. Discussion. Step A allows one to solve aproximately the problem (10.25) for $I = I(x_k, \varepsilon_k)$ confining oneself to several iterations in the direction of increase of the parameter σ. The conditions (10.27) make it possible to use the already known duplication method to calculate β_k.

The convergence of the method will be investigated

only for the case when β_k satisfies the conditions (10.26). On studying the proof of the convergence the reader, when necessary, can easily investigate the convergence for the case (10.27) as well.

10.4.6. Assumptions.

(1) the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are of class $C^{1,1}(X)$,

(2) there is a number $M > 0$ such that $\| f_i'(\mathbf{x}) \| \leqslant M$ for all $\mathbf{x} \in X_0 \quad \{\mathbf{x} \in X: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$,

(3) $X^* = \{\mathbf{x}^* \in X: \bar{\sigma}(\mathbf{x}^*, 0) \quad 0\} \neq \varnothing$,

(4) $\inf_{\mathbf{x} \in X} \varphi(\mathbf{x}) > -\infty$,

(5) the sequence $\{\mathbf{x}_k\}$ is bounded.

10.4.7. Conditions for extremality. We elucidate the condition (3). By virtue of Theorem 3.3.2, the condition that in each pair σ, $\mathbf{s}$ satisfying the system

$$\langle f_i'(\mathbf{x}), \mathbf{s}\rangle + \sigma \leqslant 0, \quad i \in I(\mathbf{x}. 0), \quad -\langle \varphi'(\mathbf{x}), \mathbf{s}\rangle + \sigma \leqslant 0,$$

there must be $\sigma \leqslant 0$ is a necessary condition for $\mathbf{x}$ to be a point of local minimum of the function $\varphi(\mathbf{x})$ on X.

If the function $\varphi(\mathbf{x})$ is convex, the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$ are concave, and, in addition, the set X is regular in Slater's sense (see Sec. 3.4.3), then the above stated conditions are sufficient for the point $\mathbf{x}$ to be optimal.

The proof of this assertion consists in the following. Repeating the calculations in the proof of Theorem 3.4.4 (also see Sec. 3.3.5) we pass to the relation

$$\varphi'(\mathbf{x}) = \sum_{i \in I(\mathbf{x}, 0)} u_i f_i'(\mathbf{x}), \quad u_i \geqslant 0, \quad i \in I(\mathbf{x}, 0),$$

which, according to Theorem 3.5.1, implies that the point $\mathbf{x}$ is optimal.

We now turn to the condition $X^* \neq \varnothing$. Since the pair $\sigma = 0$, $\mathbf{s} = 0$ is feasible for the problem (10.25), we have $\bar{\sigma}(\mathbf{x}, 0) \geqslant 0$ in its solution, and, consequently, X^* is the set of points where the necessary conditions for local minimum are fulfilled. We note that to these points a number of other points are added, e.g. those where there are no feasible directions.

10.4.8. Lemma. *For all σ and $\mathbf{s}$ satisfying the conditions*

$$\begin{gathered} \langle f_i'(\mathbf{x}), \mathbf{s}\rangle + \sigma \leqslant 0, \\ i \in I(\mathbf{x}, \varepsilon), \quad \mathbf{x} \in X, \quad \langle \mathbf{s}, \mathbf{s}\rangle \leqslant 1, \quad \sigma \geqslant \varepsilon \geqslant 0, \end{gathered} \tag{10.28}$$

the inequality

$$\zeta \geqslant C_1 \varepsilon \tag{10.29}$$

holds, where $C_1 = \min\{1/M, 1/L\}$.

Proof. It obviously suffices to consider the case $\zeta < +\infty$. Since $\mathbf{y} = \mathbf{x} - \zeta \mathbf{s}$ is a boundary point (according to the definition of ζ), there is an index i such that $f_i(\mathbf{y}) = 0$. If $f_i(\mathbf{x}) > \varepsilon$ at the point $\mathbf{x}$, then $\varepsilon < f_i(\mathbf{x}) = |f_i(\mathbf{x}) - f_i(\mathbf{y})| \leqslant M \|\mathbf{x} - \mathbf{y}\| = M\zeta$, whence

$$\zeta > \varepsilon/M. \tag{10.30}$$

Let $0 \leqslant f_i(\mathbf{x}) \leqslant \varepsilon$ at the point $\mathbf{x}$, i.e. $i \in I(\mathbf{x}, \varepsilon)$. Denote $\psi(\beta) = f_i(\mathbf{x} - \beta \mathbf{s})$ and take into consideration that $\psi(\beta) \geqslant 0$ for $\beta \in [0, \zeta]$ and $\psi(\zeta) = 0$. It can easily be seen that

$$\left.\frac{d\psi(\beta)}{d\beta}\right|_{\beta=\zeta} \leqslant 0,$$

and, consequently, $\langle f_i'(\mathbf{x}), \mathbf{s}\rangle \geqslant 0$. From $i \in I(\mathbf{x}, \varepsilon)$ and the condition that σ and $\mathbf{s}$ satisfy the system (10.28) it follows that $\langle f_i'(\mathbf{x}), \mathbf{s}\rangle \leqslant -\sigma$, and therefore

$$\varepsilon \leqslant \sigma \leqslant -\langle f_i'(\mathbf{x}), \mathbf{s}\rangle \leqslant \langle f_i'(\mathbf{y}), \mathbf{s}\rangle - \langle f_i'(\mathbf{x}), \mathbf{s}\rangle$$
$$\leqslant \|f_i'(\mathbf{y}) - f_i'(\mathbf{x})\| \|\mathbf{s}\| \leqslant L\|\mathbf{y} - \mathbf{x}\| = L\zeta,$$

whence

$$\zeta \geqslant \varepsilon/L. \tag{10.31}$$

Now, (10.30) and (10.31) imply $\zeta \geqslant \varepsilon \min\{1/M, 1/L\}$, i.e. the inequality (10.29) holds. $\triangle$

10.4.9. Lemma. *If the point* $\mathbf{x}_{k+1}$ *is constructed according to the scheme of the method and* $\sigma_k \geqslant \varepsilon_k \geqslant 0$, *then the inequality*

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant C_2 \varepsilon_k^2, \quad C_2 = \lambda C_1/2 \tag{10.32}$$

holds.

Proof. From Lemma 9.2.3, the condition (10.26), and the inequality $\langle \varphi'(\mathbf{x}_k), \mathbf{s}_k\rangle \geqslant \sigma_k$ (this inequality holds since σ_k and $\mathbf{s}_k$ are feasible for the problem (10.25)) we obtain for any $\beta \in [0, \zeta_k]$:

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1})$$
$$\geqslant \lambda_k(\varphi(\mathbf{x}_k) - \omega_k) \geqslant \lambda_k(\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta \mathbf{s}_k))$$
$$\geqslant \lambda_k\left(\beta\langle\varphi'(\mathbf{x}_k), \mathbf{s}_k\rangle - \frac{1}{2}L\beta^2\|\mathbf{s}_k\|^2\right) \geqslant \lambda_k\left(\beta\sigma_k - \frac{1}{2}L\beta^2\right).$$

For $\beta = \min\{\zeta_k, \sigma_k/L\}$ this implies the inequality

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant \frac{1}{2}\lambda_k \sigma_k \min\left\{\zeta_k \frac{\sigma_k}{L}\right\}$$

$$\geqslant \frac{1}{2}\lambda \varepsilon_k \min\left\{\zeta_k, \frac{\varepsilon_k}{L}\right\}. \quad (10.33)$$

By virtue of Lemma 10.4.8, we arrive at (10.32) for $C_2 = \lambda C_1/2$. $\triangle$

10.4.10. Lemma. *For any point* $\bar{\mathbf{x}} \in X$ *there are* $\bar{\varepsilon} = \bar{\varepsilon}(\bar{\mathbf{x}}) > 0$ *and* $\delta = \delta(\bar{\varepsilon}) > 0$ *such that* $I(\mathbf{x}, \varepsilon) \subset I(\bar{\mathbf{x}}, 0)$ *for all* $\varepsilon \in [0, \bar{\varepsilon}]$ *and* $\mathbf{x} \in U_\delta(\bar{\mathbf{x}}) = \{\mathbf{x} \in X : \| \mathbf{x} - \bar{\mathbf{x}} \| \leqslant \delta\}$.

Proof. Denote $J = \{i = 1, 2, \ldots, m\}$ and take $\bar{\varepsilon} = \min\limits_{i \in J \setminus I(\bar{\mathbf{x}}, 0)} \{f_i(\bar{\mathbf{x}})\}/2$. We note that $\bar{\varepsilon} > 0$ since $f_i(\bar{\mathbf{x}}) > 0$ for $i \in J \setminus I(\bar{\mathbf{x}}, 0)$.

Since the functions $f_i(\mathbf{x})$ are continuous and the number of indices i, $i \leqslant m$, is finite, there is $\delta(\bar{\varepsilon}) > 0$ such that $|f_i(\bar{\mathbf{x}}) - f_i(\mathbf{x})| \leqslant \bar{\varepsilon}$ for all $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$ and all $i \in J$. Since $f_i(\bar{\mathbf{x}}) \geqslant 2\bar{\varepsilon}$, $i \in J \setminus I(\bar{\mathbf{x}}, 0)$, we have $f_i(\mathbf{x}) \geqslant \varepsilon$ for all $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$ and all $i \in J \setminus I(\bar{\mathbf{x}}, 0)$. Hence, for $\varepsilon \in [0, \bar{\varepsilon}]$ we have $i \in J \setminus I(\mathbf{x}, \varepsilon)$ for any $i \in J \setminus I(\bar{\mathbf{x}}, 0)$, and, consequently, the inclusion $I(\mathbf{x}, \varepsilon) \subset I(\bar{\mathbf{x}}, 0)$ takes place. $\triangle$

Let $\bar{\sigma} = \bar{\sigma}(\bar{\mathbf{x}}, 0)$, $\bar{\mathbf{s}} = \bar{\mathbf{s}}(\bar{\mathbf{x}}, 0)$ be the solution to the problem (10.25) for $\mathbf{x} = \bar{\mathbf{x}}$ and $I = I(\bar{\mathbf{x}}, 0)$.

10.4.11. Lemma. *If* $I(\mathbf{x}, \varepsilon) \subset I(\bar{\mathbf{x}}, 0)$ *and* $\bar{\sigma} > 0$, *then there is* $\delta = \delta(\bar{\sigma}) > 0$ *such that* $\sigma(\mathbf{x}, \varepsilon) \geqslant \bar{\sigma}/2$ *for all* $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$.

Proof. Since $\varphi'(\mathbf{x})$ and $f_i'(\mathbf{x})$ $(i = 1, 2, \ldots, m)$ are continuous, there is $\delta(\bar{\sigma}) > 0$ such that for all $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$ we have $\| \varphi'(\bar{\mathbf{x}}) - \varphi'(\mathbf{x}) \| \leqslant \bar{\sigma}/2$ and $\| f_i'(\mathbf{x}) - f_i'(\bar{\mathbf{x}}) \| \leqslant \bar{\sigma}/2$, $i \in I(\bar{\mathbf{x}}, 0)$. Further, for any $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$ and any $i \in I(\mathbf{x}, \varepsilon) \subset I(\bar{\mathbf{x}}, 0)$ there holds

$$0 \geqslant \langle f_i'(\bar{\mathbf{x}}), \bar{\mathbf{s}} \rangle + \bar{\sigma} = \langle f_i'(\mathbf{x}), \bar{\mathbf{s}} \rangle + \bar{\sigma} + \langle f_i'(\bar{\mathbf{x}}) - f_i'(\mathbf{x}), \bar{\mathbf{s}} \rangle$$

$$\geqslant \langle f_i'(\mathbf{x}) \bar{\mathbf{s}} \rangle + \bar{\sigma} - \| f_i'(\bar{\mathbf{x}}) - f_i'(\mathbf{x}) \| \| \bar{\mathbf{s}} \| \geqslant \langle f_i'(\mathbf{x}), \bar{\mathbf{s}} \rangle + \bar{\sigma}/2,$$

and, similarly, $0 \geqslant -\langle \varphi'(\mathbf{x}), \bar{\mathbf{s}} \rangle + \bar{\sigma}/2$.

Thus, for any $\mathbf{x} \in U_\delta(\bar{\mathbf{x}})$, $\sigma = \bar{\sigma}/2$ and $\mathbf{s} = \bar{\mathbf{s}}$ satisfy the constraints in the problem (10.25) for $I = I(\mathbf{x}, \varepsilon)$.

However, since $\sigma(x, \varepsilon)$ and $s(x, \varepsilon)$ is a solution to the problem of maximizing the parameter σ under the same constraints, we have $\sigma(x, \varepsilon) \geqslant \bar{\sigma}/2$. $\triangle$

10.4.12. Theorem. *If the assumptions* (1)-(5) *hold, then*

$$\lim_{k \to \infty} \rho(x_k, X^*) = \lim_{k \to \infty} \inf_{x^* \in X^*} \| x_k - x^* \| = 0.$$

Proof. Denote by K the set of all indices in the sequence $\{x_k\}$: $K = \{k = 0, 1, \ldots\}$. Assume that $\varepsilon_k \geqslant \varepsilon > 0$ for some $\varepsilon > 0$ and all $k \in K$. Then, according to the scheme of the method, there is an index $k_0 \in K$ such that $\sigma_k \geqslant \varepsilon_k \geqslant \varepsilon > 0$ for all $k \geqslant k_0$. From (10.29) we obtain $\zeta_k \geqslant C_1 \varepsilon_k \geqslant C_1 \varepsilon$, and therefore, by virtue of (10.32), we have $\varphi(x_k) - \varphi(x_{k+1}) \geqslant C_2 \varepsilon^2$ for all $k \geqslant k_0$. However, the sequence $\{\varphi(x_k)\}$ is convergent (since it is monotonic and bounded), which contradicts the foregoing inequality. Hence, $\varepsilon_k \to 0$, $k \to \infty$. By the scheme of the method, there is a set of indices $K_1 \subset K$ such that $\sigma_k \to 0$, $k \in K_1$, $k \to \infty$.

We now prove that all the limiting points of the sequence $\{x_k\}$ belong to the set X^*. We consider the following two cases.

Case 1. Let the sequence $\{x_k\}$ have a single limiting point $\bar{x}$, i.e. $\lim_{k \to \infty} x_k = \bar{x} \in X$. Assume that $\bar{x} \in X \setminus X^*$. Then $\bar{\sigma}(\bar{x}, 0) = \bar{\sigma} > 0$, and, consequently, by Lemma 10.4.10, there are $\bar{\varepsilon} > 0$ and $\delta(\bar{\varepsilon}) > 0$ such that $I(x, \varepsilon) \subset I(\bar{x}, 0)$ for all $\varepsilon \in [0, \bar{\varepsilon}]$ and $x \in U_\delta(\bar{x})$. Since $\varepsilon_k \to 0$ and $x_k \to \bar{x}$, $k \to \infty$, there is an index k_0 such that $x_k \in U_\delta(\bar{x})$ and $\varepsilon_k < \bar{\varepsilon}$ for all $k \geqslant k_0$, and hence $I(x_k, \varepsilon_k) \subset I(\bar{x}, 0)$. However, according to Lemma 10.4.11, for $k \geqslant k_0$ there holds $\bar{\sigma}(x_k, \varepsilon_k) \geqslant \bar{\sigma}/2$, which contradicts the fact that $\sigma_k \to 0$, $k \in K_1$, $k \to \infty$.

Case 2. Let there exist a limiting point $\tilde{x} \in X$ of the sequence $\{x_k\}$ distinct from $\bar{x}$: $\tilde{x} \neq \bar{x}$. As before, we assume that $\bar{x} \in X \setminus X^*$. Then there is $\delta > 0$ such that $\tilde{x} \in X \setminus U_\delta(\bar{x})$. Since $\bar{x}$ and $\tilde{x}$ are two different limiting points of the sequence, for any natural N there is an index $k \geqslant N$ and a number $m \geqslant 1$ such that $x_k \in U_{\delta/2}(\bar{x})$, $x_{k+i} \in U_\delta(\bar{x})$, $i = 0, 1, \ldots, m-1$; $x_{k+m} \in X \setminus U_\delta(\bar{x})$.

We now make use of the inequality (10.33):

$$\varphi(x_k) - \varphi(x_{k+m}) = \sum_{i=k}^{k+m-1} \varphi(x_i) - \varphi(x_{i+1}) \geqslant \frac{1}{2}\alpha\lambda \sum_{i=k}^{k+m-1} \sigma_i \min\left\{\zeta_i, \frac{\sigma_i}{L}\right\},$$

and, since $x_i \in U_\delta(\bar{x})$, $i = k, k+1, \ldots, k+m-1$, we have $\sigma_i \geqslant \bar{\sigma}/2$, $i = k, k+1, \ldots, k+m-1$. Moreover, $\zeta_i \geqslant \beta_i$ for all indices i, and therefore

$$\varphi(x_k) - \varphi(x_{k+m}) \geqslant \frac{1}{4}\alpha\lambda\bar{\sigma} \sum_{i=k}^{k+m-1} \min\left\{\beta_i, \frac{\bar{\sigma}}{2L}\right\}.$$

We note that

$$\Sigma \triangleq \sum_{i=k}^{k+m-1} \min\left\{\beta_i, \frac{\bar{\sigma}}{2L}\right\} \geqslant \min\left\{\sum_{i=k}^{k+m-1} \beta_i, \frac{\bar{\sigma}}{2L}\right\}.$$

Indeed, if $\beta_i \leqslant \bar{\sigma}/(2L)$ for all i, then $\Sigma = \sum_{i=1}^{k+m-1} \beta_i$. In case $\beta_j > \bar{\sigma}/(2L)$ for at least one index j, we have $\Sigma > \bar{\sigma}/(2L)$. Since $\| x_{k+m} - x_k \| \geqslant \delta/2$, there must be $\sum_{i=k}^{k+m-1} \beta_i \geqslant \delta/2$, and therefore

$$\varphi(x_k) - \varphi(x_{k+m}) \geqslant \frac{1}{4}\alpha\lambda\bar{\sigma} \min\left\{\frac{\delta}{2}, \frac{\bar{\sigma}}{2L}\right\} = \text{const} > 0.$$

The last inequality contradicts the convergence of the sequence $\{\varphi(x_k)\}$.

Thus, any limiting point of the sequence $\{x_k\}$ belongs to the set X^*. In view of the compactness of $\{x_k\}$, it follows that $\lim_{k\to\infty} \rho(x_k, X^*) = 0$. $\triangle$

10.4.13. The case involving linear constraints. If some of the constraints determining the set X are linear, it is advisable to split the constraints into two groups, linear and nonlinear:

$$X = \{x: f_i(x) \geqslant 0, \quad i \in I_1, \quad \langle a_i, x\rangle \geqslant b_i, \quad i \in I_2\}, \quad I_1 \cup I_2 = \{1, 2, \ldots, m\}. \tag{10.34}$$

In this case the problem of determining the direction of descent takes the following form:

$$\begin{gathered} \max \sigma, \\ \langle f_i'(x_k),\ s\rangle + \sigma \leqslant 0, \quad i \in I_1(x_k,\ \varepsilon_k), \\ \langle a_i,\ s\rangle \leqslant 0, \quad i \in I_2(x_k,\ \varepsilon_k), \\ -\langle \varphi'(x_k,\ s)\rangle + \sigma \leqslant 0, \\ \langle s,\ s\rangle \leqslant 1, \end{gathered} \tag{10.35}$$

where

$$I_1(x_k,\ \varepsilon_k) = \{i \in I_1:\ 0 \leqslant f_i(x_k) \leqslant \varepsilon_k\},$$
$$I_2(x_k,\ \varepsilon_k) = \{i \in I_2:\ 0 \leqslant \langle a_i,\ x_k\rangle - b_i \leqslant \varepsilon_k\}.$$

Naturally, the scheme of the method remains the same in this case. The argument in the proof of the convergence of the method also remains the same. However, here the condition of regularity can be weakened as compared to the one introduced earlier. Namely, the set X determined by (10.34) will be called regular if for all $i \in I_1$ there is $x \in X$ such that $f_i(x) > 0$ $(i \in I_1)$. Thus, the former property of regularity relates only to the nonlinear constraints.

10.4.14. On the solution of the problem of determining the direction of descent. When the last constraint in the problem (10.35) has the form $-1 \leqslant s_j \leqslant 1$, $j = \overline{1, n}$, the choice of the direction of descent is carried out by solving a linear programming problem admitting of the application of finite algorithms. In case $\langle s, s\rangle \leqslant 1$, the problem (10.25) involves a nonlinear constraint. We shall show that, in a sense, this problem is equivalent to a quadratic programming problem of a special type.

We write the problem (10.25) thus:

$$\begin{gathered} \max \sigma, \\ Qs + \sigma I \leqslant 0, \\ \langle s,\ s\rangle \leqslant 1. \end{gathered} \tag{10.36}$$

Here Q is a matrix whose rows are the vectors $[f_i'(x)]^T$ for $i \in I(x, \varepsilon)$ and $[-\varphi'(x)]^T$. The index k indicating the number of the iteration is omitted here to make the expression less cumbersome. The symbol I denotes the vector $I^T = (1, 1, \ldots, 1)$.

It should be noted that since $\sigma = 0$, $\mathbf{s} = 0$ is a feasible pair and the feasible region in the problem (10.36) is bounded and closed, the problem is always solvable.

Consider the problem

$$\begin{aligned} &\min \langle \mathbf{s}, \mathbf{s} \rangle, \\ &Q\mathbf{s} + \mathbf{I} \leqslant \mathbf{0}. \end{aligned} \tag{10.37}$$

It is obvious that if the set $\{\mathbf{s}: Q\mathbf{s} + \mathbf{I} \leqslant \mathbf{0}\}$ is nonempty, the solution to this problem exists. We denote it as $\mathbf{s}_1$. We note that $\mathbf{s}_1 \neq \mathbf{0}$ since $\mathbf{s} = \mathbf{0}$ is not a feasible point of the problem.

10.4.15. Theorem. (1) *The pair* $\sigma^* = 1/\| \mathbf{s}_1 \|$, $\mathbf{s}^* = \mathbf{s}_1/\| \mathbf{s}_1 \|$ *is a solution to the problem* (10.36).

(2) *If the set* $\{\mathbf{s}: Q\mathbf{s} + \mathbf{I} \leqslant \mathbf{0}\}$ *is empty, then the optimal value of* σ *in the problem* (10.36) *is equal to zero.*

Proof. Since

$$Q\mathbf{s}^* + \sigma^* \mathbf{I} = (Q\mathbf{s}_1 + \mathbf{I})/\| \mathbf{s}_1 \| \leqslant 0 \quad \text{and} \quad \langle \mathbf{s}^*, \mathbf{s}^* \rangle = 1,$$

the pair σ^*, $\mathbf{s}^*$ is feasible for the problem (10.36). Assume that there is a pair $\bar{\sigma}$, $\bar{\mathbf{s}}$ satisfying the constraints of the problem (10.36) such that $\bar{\sigma} > \sigma^* > 0$. Then $\bar{\mathbf{s}}/\bar{\sigma}$ is a feasible point for the problem (10.37)

$$Q\frac{1}{\bar{\sigma}}\bar{\mathbf{s}} + \mathbf{I} = \frac{1}{\bar{\sigma}}(Q\bar{\mathbf{s}} + \bar{\sigma}\mathbf{I}) \leqslant 0,$$

and, consequently,

$$\begin{aligned} \langle \mathbf{s}_1, \mathbf{s}_1 \rangle &\leqslant \left\langle \frac{1}{\bar{\sigma}}\bar{\mathbf{s}}, \frac{1}{\bar{\sigma}}\bar{\mathbf{s}} \right\rangle \\ &= \frac{1}{\bar{\sigma}^2}\langle \bar{\mathbf{s}}, \bar{\mathbf{s}} \rangle < \left(\frac{1}{\sigma^*}\right)^2 \langle \bar{\mathbf{s}}, \bar{\mathbf{s}} \rangle = \langle \mathbf{s}_1, \mathbf{s}_1 \rangle \langle \bar{\mathbf{s}}, \bar{\mathbf{s}} \rangle, \end{aligned}$$

whence we obtain the inequality $\langle \bar{\mathbf{s}}, \bar{\mathbf{s}} \rangle > 1$ contradicting the fact that $\bar{\mathbf{s}}$ is a feasible point for (10.36). The second part of the theorem is an obvious consequence of its first part. $\triangle$

10.4.16. To solve the problem (10.37) it is convenient to resort to the following considerations. The expression

$$L(\mathbf{s}, \mathbf{u}) = \| \mathbf{s} \|^2 + \langle \mathbf{u}, Q\mathbf{s} + \mathbf{I} \rangle, \quad \mathbf{s} \in E_n, \quad \mathbf{u} \geqslant \mathbf{0},$$

is the Lagrange function for the problem (10.37). Its sad-

dle point $\mathbf{s}^*$, $\mathbf{u}^*$ is determined by the condition

$$L(\mathbf{s}^*, \mathbf{u}^*) = \min_{\mathbf{s}} \max_{\mathbf{u} \geqslant 0} L(\mathbf{s}, \mathbf{u}).$$

Since the function $L(\mathbf{s}, \mathbf{u})$ is strongly convex and quadratic in $\mathbf{s}$, its minimum with respect to $\mathbf{s}$ is attained at the point $\mathbf{s}(\mathbf{u}) = -Q^T\mathbf{u}/2$. Therefore $L(\mathbf{s}(\mathbf{u}), \mathbf{u})$ takes the form $L(\mathbf{s}(\mathbf{u}), \mathbf{u}) = -\| Q^T \mathbf{u} \|^2/4 + \langle \mathbf{I}, \mathbf{u} \rangle$. Hence, to find $\mathbf{s}^*$, one must solve the problem

$$\max_{\mathbf{u} \geqslant 0} L(\mathbf{s}(\mathbf{u}), \mathbf{u}).$$

This is the problem of minimizing a quadratic function on the set $\mathbf{u} \geqslant \mathbf{0}$ having a simple structure. The problem can be appropriately solved by the method of projecting conjugate gradients. This method differs from the earlier presented gradient projection method in that one projects on the set $\mathbf{u} \geqslant \mathbf{0}$, not the gradients of the minimized function but the conjugate gradients.

10.5. Methods of Finding a Feasible Point

10.5.1. In all the relaxation methods for solving mathematical programming problems considered earlier it was assumed that an initial point $\mathbf{x}_0 \in X$ was known. Presented below are three methods of finding a point $\mathbf{x}$ belonging to the set

$$X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0,\ i \in I_1,\ \langle \mathbf{a}_i, \mathbf{x} \rangle \geqslant b_i,\ i \in I_2\}.$$

To determine a point $\mathbf{x}$ satisfying the linear constraints $\langle \mathbf{a}_i, \mathbf{x} \rangle \geqslant b_i$, $i \in I_2$, any of the methods considered in linear programming (see the methods of finding an initial extreme point) applies. We note that these methods are finite.

Thus, we shall assume that we know a point $\mathbf{x}_0 \in E_n$ such that $\langle \mathbf{a}_i, \mathbf{x}_0 \rangle \geqslant b_i$, $i \in I_2$. Denote

$$h_i(\mathbf{x}) = \begin{cases} -f_i(\mathbf{x}), & i \in I_1, \\ b_i - \langle \mathbf{a}_i, \mathbf{x} \rangle, & i \in I_2. \end{cases}$$

10.5.2. First Method. Let

$$\rho_i = \begin{cases} 0 \text{ if } h_i(\mathbf{x}_0) \leqslant 0, & i \in I_1, \\ 1 \text{ if } h_i(\mathbf{x}_0) > 0, & i \in I_2. \end{cases}$$

To find a feasible point of the set X we solve the following problem:

$$\begin{gathered}\xi \to \inf, \\ f_i(\mathbf{x}) + \rho_i \xi \geqslant 0, \quad i \in I_1, \\ \langle \mathbf{a}_i, \mathbf{x} \rangle \geqslant b_i, \quad i \in I_2.\end{gathered} \tag{10.38}$$

As an initial feasible pair $\mathbf{x}_0$, ξ_0 for this problem we take $\mathbf{x}_0$ and $\xi_0 = \max\limits_{i \in I_1, \rho_i = 1} h_i(\mathbf{x}_0)$. Denote by $\mathbf{x}^*$, ξ^* the solution to the problem (10.38). We note that, generally, since $\xi^* = \inf \xi$, this notation includes the case $\xi^* = -\infty$ as well.

10.5.3. *If for the problem* (10.38) *there is a feasible pair* $\mathbf{x}$, ξ *such that* $\xi \leqslant 0$, *then the point* $\mathbf{x}$ *belongs to the set* X. *In case* $\xi^* > 0$ *in the solution* $\mathbf{x}^*$, ξ^* *of the problem* (10.38), *the set* X *is empty.* Indeed, if $X \neq \varnothing$, then there is a point $\mathbf{x}$ such that $h_i(\mathbf{x}) \leqslant 0$, for all $i \in I_1 \cup I_2$. In this case, taking $\xi = \max\limits_{i \in I_1} h_i(\mathbf{x}) \leqslant 0$, we see that $\mathbf{x}$, ξ is a feasible pair, and the inequality $\xi < \xi^*$ holds, which contradicts the assumption that ξ^* is the minimum value. $\triangle$

The problem (10.38) can be solved using any convergent relaxation method.

10.5.4. Theorem. *If the set* X *is nonempty and possesses the property of regularity relative to the nonlinear constraints, then any weakly convergent relaxation process yields a feasible pair* $\mathbf{x}$, $\xi \leqslant 0$ *of the problem* (10.38) *after a finite number of iterations.*

Proof. If $\xi_0 \leqslant 0$, then $\mathbf{x}_0 \in X$. Assume that $\xi_0 > 0$. The set X being regular, there is a point $\bar{\mathbf{x}} \in X$ such that $h_i(\bar{\mathbf{x}}) < 0$ for all $i \in I_1$ and $\langle \mathbf{a}_i, \bar{\mathbf{x}} \rangle \geqslant b_i$ for all $i \in I_2$. It follows that the pair $\bar{\mathbf{x}}$, $\bar{\xi} = \max\limits_{i \in I_1, \rho_i = 1} h_i(\bar{\mathbf{x}})$ satisfies the constraints in the problem (10.38), and hence $\lim\limits_{m \to \infty} \xi_m = \xi^* \leqslant \bar{\xi} < 0$. However, the initial value ξ_0 is positive ($\xi_0 > 0$), and therefore the properties of the convergent sequence $\{\xi_m\}$ imply that there is an index m_0 beginning with which $\xi_m \leqslant 0$, $m = m_0, m_0 + 1, \ldots$.

Thus, any weakly convergent relaxation process yields a pair $\mathbf{x}_{m_0}$, $\xi_{m_0} \leqslant 0$ after a finite number m_0 of iterations. $\triangle$

10.5.5. Second method. Let $\mathbf{x}_0$ be a fixed point in E_n. Denote

$$\rho_i = \begin{cases} 0 \text{ if } h_i(\mathbf{x}_0) \leqslant 0, & i \in I_1 \cup I_2, \\ 1 \text{ if } h_i(\mathbf{x}_0) > 0, & i \in I_1 \cup I_2. \end{cases}$$

To find a feasible point in the region X we shall solve the following problem:

$$\begin{aligned} \xi &\to \min, \\ f_i(\mathbf{x}) + \rho_i \xi &\geqslant 0, \quad i \in I_1, \\ \langle \mathbf{a}_i, \mathbf{x} \rangle + \rho_i \xi &\geqslant b_i, \quad i \in I_2. \end{aligned} \tag{10.39}$$

As an initial feasible pair $\mathbf{x}_0$, ξ_0 for this problem one takes $\mathbf{x}_0$, $\xi_0 = \max\limits_{i,\ \rho_i = 1} h_i(\mathbf{x}_0)$.

It is obvious that if there is a feasible pair $\mathbf{x}$, ξ such that $\xi = 0$, then the point $\mathbf{x}$ belongs to the set X. In case $\xi^* > 0$ in the solution $\mathbf{x}^*$, ξ^* to the problem (10.39), the set X is empty.

In contrast to the foregoing method, here a feasible point of the set X is found in one stage (it is unnecessary to solve separately the problem of finding a point satisfying the linear constraints); at the same time, the finiteness of the relaxation process in the second method can only be guaranteed when $\xi^* < 0$, for instance, when the set X is regular relative to both the linear and nonlinear constraints.

10.5.6. Third method. We consider the basic convex programming problem $\varphi(\mathbf{x}) \to \min$ under the conditions

$$X = \{\mathbf{x} \in \Gamma: f_i(\mathbf{x}) \geqslant 0, \ i = \overline{1, m}\}, \quad \Gamma = \{\mathbf{x}: \mathbf{x} \geqslant 0\}.$$

The third method is a generalization of the M-method for solving linear programming problems to the nonlinear (convex) case.

Given a fixed point $\mathbf{x}_0 > 0$, we introduce

$$\rho_i = \begin{cases} 0 & \text{if } f_i(\mathbf{x}_0) \geqslant 0, \\ 1 & \text{if } f_i(\mathbf{x}_0) < 0, \end{cases}$$

and consider the problem

$$\varphi(\mathbf{x}) + M\xi \to \min,$$
$$f_i(\mathbf{x}) + \rho_i \xi \geqslant 0, \quad i = \overline{1, m}, \tag{10.40}$$
$$\mathbf{x} \geqslant \mathbf{0}, \quad \xi \geqslant 0.$$

Here M is a number whose choice will be specified later.

The pair $\mathbf{x}_0$, $\xi_0 = \max\limits_{i,\ \rho_i = 1} (-f_i(x_0))$ is feasible for the problem (10.40). We shall prove that under certain conditions the solution $\mathbf{x}^*$, ξ^* to this problem is such that $\mathbf{x}^*$ is the optimal point of the basic problem.

Let all the conditions be fulfilled under which Theorem 3.6.7 holds, namely, let the set X in the basic convex programming problem possess the regularity property 3.4.2 and let the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ $(i = \overline{1, m})$ be continuously differentiable on the set Γ. Under these assumptions the following theorem holds.

10.5.7. Theorem. *If the basic convex programming problem is solvable, then there is a number M_0 such that for all $M \geqslant M_0$ the point $\mathbf{x}^*$ involved in any solution $\mathbf{x}^*$, ξ^* to the problem* (10.40) *is optimal for the basic problem.*

Proof. By the statement of the theorem, there is a point

$$\mathbf{x}^* = \arg\min \{\varphi(\mathbf{x}): \mathbf{x} \in X\}.$$

The expression $L(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle$, where $\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$, is the Lagrange function for the basic convex programming problem. By virtue of Theorem 3.6.7, the pair $\mathbf{x}^*$, $\mathbf{y}^*$ is a saddle point of the function $L(\mathbf{x}, \mathbf{y})$ on the set $\mathbf{x} \geqslant \mathbf{0}$, $\mathbf{y} \geqslant \mathbf{0}$, and $L(\mathbf{x}^*, \mathbf{y}^*) = \varphi^*$ since $\langle \mathbf{y}^*, \mathbf{f}(\mathbf{x}^*) \rangle = 0$ (see (3.34) in Sec. 3.6.7). We denote $\boldsymbol{\rho}^T = (\rho_1, \ldots, \rho_m)$ and write down the Lagrange function for the problem (10.40):

$$L_1(\mathbf{x}, \xi, \mathbf{y}) = \varphi(\mathbf{x}) + M\xi - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) + \boldsymbol{\rho}\xi \rangle,$$

$$\inf_{\mathbf{x} \geqslant \mathbf{0},\ \xi \geqslant 0} L_1(\mathbf{x}, \xi, \mathbf{y})$$
$$= \inf_{\mathbf{x} \geqslant \mathbf{0},\ \xi \geqslant 0} (\varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle + \xi (M - \langle \mathbf{y}, \boldsymbol{\rho} \rangle))$$
$$= \begin{cases} \inf\limits_{\mathbf{x} \geqslant \mathbf{0}} (\varphi(\mathbf{x}) - \langle \mathbf{y}, \mathbf{f}(\mathbf{x}) \rangle) & \text{for } M - \langle \mathbf{y}, \boldsymbol{\rho} \rangle \geqslant 0, \\ -\infty & \text{if otherwise.} \end{cases}$$

Consequently, if $M \geqslant \sum_{i=1}^{m} y_i^* \equiv M_0$, then

$$\inf_{x \geqslant 0,\ \xi \geqslant 0} L_1(x, \xi, y^*) = \inf_{x \geqslant 0} L(x, y^*) = L(x^*, y^*) = \varphi^*.$$

On the other hand,

$$\inf_{x \geqslant 0,\ \xi \geqslant 0} \sup_{y \geqslant 0} L_1(x, \xi, y) \geqslant \sup_{y \geqslant 0} \inf_{x \geqslant 0,\ \xi \geqslant 0} L_1(x, \xi, y) \geqslant \inf_{x \geqslant 0,\ \xi \geqslant 0} L_1(x, \xi, y^*) = \varphi^*.$$

Denote $Z = \{x, \xi: f(x) + \rho\xi \geqslant 0, \quad x \geqslant 0, \quad \xi \geqslant 0\}$. Since $L_1(x, \xi, y)$ is the Lagrange function for the problem (10.40), we have

$$\inf_{x \geqslant 0,\ \xi \geqslant 0} \sup_{y \geqslant 0} L_1(x, \xi, y) = \inf_{(x, \xi) \in Z} (\varphi(x) + M\xi) \leqslant \varphi^*,$$

where the right-hand inequality holds owing to the fact that $(x, 0) \in Z$ for any $x \in X$. Thus,

$$\inf_{x \geqslant 0,\ \xi \geqslant 0} \sup_{y \geqslant 0} L_1(x, \xi, y) = \sup_{y \geqslant 0} \inf_{x \geqslant 0,\ \xi \geqslant 0} L_1(x, \xi, y) = L_1(x^*, 0, y^*) = L(x^*, y^*) = \varphi^*,$$

which implies the assertion of the theorem. △

10.5.8. Corollary. *If there is a number M so large that $\xi^* > 0$ in the solution x^*, ξ^* to the M-problem* (10.40), *then the basic convex programming problem is unsolvable.*

Indeed, if the basic problem is solvable, then for $M > M_0$ we have $\xi^* = 0$ in any solution to the M-problem. This contradiction proves the corollary. △

10.5.9. Discussion. 1. In practice, the magnitude of M_0 is of course unknown. Therefore one takes as M a sufficiently large number. Here one should take into account that the choice of an excessively large M may result in a loss of calculation accuracy.

2. If the linear constraints are separated out from the conditions determining the set X, i.e.

$$X = \{x: f_i(x) \geqslant 0;\ i \in I_1,\ \langle a_i, x \rangle \geqslant b_i,\ i \in I_2,\ x \geqslant 0\},$$

then Theorem 10.5.7 remains true when the set X is regular only relative to the nonlinear constraints.

3. Finally, we indicate a possibility of applying the M-method. Assume that a minimization process yields a point x_0 which is close to x^* but does not belong to the set X. In this case the application of the third method may prove very suitable for making the approximate solution x_0 more precise.

10.6. Methods of Random Descent

10.6.1. In these methods one takes a realization of an n-dimensional random vector s with a known probability distribution as the direction s_k in the formula $x_{k+1} = x_k - \beta_k s_k$ ($k = 0, 1, \ldots$). It is difficult to give a definite characteristic of the effectiveness of the methods of random descent; however, owing to the application of high-speed computers they prove practically useful.

10.6.2. Notation and definitions. $P\{A \mid \xi_{k-1}, \ldots, \xi_0\}$ is the conditional probability of an event A relative to random variables $\xi_{k-1}, \ldots, \xi_0$.

$M(\xi_k \mid \xi_{k-1}, \ldots, \xi_0)$ is the conditional mathematical expectation of a random variable ξ_k relative to random variables $\xi_{k-1}, \ldots, \xi_0$.

The expression $\xi_k \xrightarrow[P]{} \xi$, $k \to \infty$, means that the sequence of random variables $\{\xi_k\}$ converges in probability to a random variable ξ, i.e. given any $\varepsilon > 0$, we have $P\{|\xi_k - \xi| > \varepsilon\} \to 0$, $k \to \infty$.

The expression $\xi_k \xrightarrow[\text{a.c.}]{} \xi$, $k \to \infty$, symbolizes that the sequence of random variables $\{\xi_k\}$ converges almost certainly (with probability one) to a random variable ξ, i.e. for any $\varepsilon > 0$ we have $P\{\sup_{m \geqslant k} |\xi_m - \xi| > \varepsilon\} \to 0$, $k \to \infty$.

10.6.3. Random search in the solution of unconditional minimization problems. One takes a random vector of unit length as s_k and then, after a small step in the direction $-s_k$ is made, verifies whether the function $\varphi(x)$ decreases in the neighborhood of the point x_k along this direction. This means in fact that the sign of the expression $\alpha_k = \langle \varphi'(x_k), s_k \rangle / \| \varphi'(x_k) \|$ is checked (we remind the reader that $\| s_k \| = 1$).

If $\alpha_k > 0$, the descent is performed according to the formula

$$x_{k+1} = x_k - \beta_k s_k, \tag{10.41}$$

where β_k is chosen so that the condition

$$\varphi(x_k - \beta_k s_k) \leqslant (1 - \lambda_k)\varphi(x_k) + \lambda_k \omega_k,$$
$$0 < \lambda \leqslant \lambda_k \leqslant 1, \quad \omega_k = \inf_{\beta \geqslant 0} \varphi(x_k - \beta s_k), \tag{10.42}$$

is fulfilled. If $\alpha_k < 0$, one puts $x_{k+1} = x_k$, $-s_{k+1} = s_k$ and performs the descent. For $\alpha_k = 0$ one puts $x_{k+1} = x_k$ and calculates a new random direction.

10.6.4. Assumptions. The random vectors $s_0, s_1, \ldots$ are assumed to be independent and such that the condition

$$P\{\alpha_k^2 \geqslant \alpha^2 \mid x_k, \ldots, x_0\} \geqslant p > 0, \quad k = 0, 1, \ldots \tag{10.43}$$

holds for some $\alpha \in (0, 1)$.

It should be noted that, generally, the random variables α_k^2 are not independent.

Examples of random vectors s_k satisfying the conditions (10.43) are considered in Sec. 10.6.10.

It is assumed that the set of stationary points of the minimized function $\varphi(x)$ is nonempty: $X^* = \{x^*: \varphi'(x) = 0\} \neq \varnothing$.

10.6.5. Theorem. *If*

(a) $\varphi(x) \in C^{1,1}(E_n)$,

(b) $\inf \varphi(x) > -\infty$,

(c) *for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that* $\|\varphi'(x)\| \geqslant \delta$ *for all* x *such that* $\rho(x, X^*) \geqslant \varepsilon$, *then* $\rho(x_k, X^*) \xrightarrow[P]{} 0$, $k \to \infty$.

Proof. By virtue of Lemma 9.4.8, we have the inequality $\varphi(x_k) - \varphi(x_{k+1}) \geqslant h\alpha_k^2 \|\varphi'(x_k)\|^2$ $(k = 0, 1, \ldots)$ for $h = \lambda/(2L)$. Since for fixed $x_k, \ldots, x_0$ the expression $\|\varphi'(x_k)\|^2$ is a constant that can be taken out of the sign of conditional mathematical expectation, in view of conditions (10.43), the above inequality implies

$$M(\varphi(x_k) - \varphi(x_{k+1}) \mid x_k, \ldots, x_0)$$
$$\geqslant h \|\varphi'(x_k)\|^2 M(\alpha_k^2 \mid x_k, \ldots, x_0)$$
$$\geqslant h \|\varphi'(x_k)\|^2 p\alpha^2, \quad k = 0, 1, \ldots.$$

As is known from probability theory, $M\xi = M(M\xi \mid \eta)$ where $M(\xi \mid \eta)$ is the conditional mathematical expecta-

tion of the random variable ξ relative to the random variable η. In view of the last inequality, we obtain, for $j = 1, k$,

$$M(\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_0)$$
$$= M(M(\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_1, \mathbf{x}_0) \mid \mathbf{x}_0)$$
$$= \ldots = M(\ldots M(M(\varphi(\mathbf{x}_j)$$
$$- \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_j, \ldots, \mathbf{x}_0) \mid \mathbf{x}_{j-1}, \ldots, \mathbf{x}_0) \ldots) \mid \mathbf{x}_0)$$
$$\geqslant h p \alpha^2 M(\ldots M(M(\| \varphi'(\mathbf{x}_j) \|^2 \mid \mathbf{x}_{j-1}, \ldots, \mathbf{x}_0) \mid \mathbf{x}_{j-2}, \ldots$$
$$\ldots, \mathbf{x}_0) \ldots \mid \mathbf{x}_0) = h p \alpha^2 M(\| \varphi'(\mathbf{x}_j) \|^2 \mid \mathbf{x}_0).$$

Therefore,

$$M(\varphi(\mathbf{x}_0) - \varphi(\mathbf{x}_{k+1}) \mid \mathbf{x}_0)$$
$$= M\left(\sum_{j=0}^{k} \varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_0\right) = \sum_{j=0}^{k} M(\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_0)$$
$$= M(\varphi(\mathbf{x}_0) - \varphi(\mathbf{x}_1) \mid \mathbf{x}_0) + \sum_{j=1}^{k} M(\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \mid \mathbf{x}_0)$$
$$\geqslant h p \alpha^2 \| \varphi'(\mathbf{x}_0) \|^2 + h p \alpha^2 \sum_{j=1}^{k} M(\| \varphi'(\mathbf{x}_j) \|^2 \mid \mathbf{x}_0),$$
$$k = 1, 2, \ldots$$

Since $\mathbf{x}_0$ is a fixed point, all the conditional mathematical expectations involved coincide with unconditional ones, and hence $M\varphi(\mathbf{x}_{k+1})$ satisfies the inequality

$$M\varphi(\mathbf{x}_{k+1}) \leqslant \varphi(\mathbf{x}_0) - h p \alpha^2 \sum_{j=0}^{k} M \| \varphi'(\mathbf{x}_j) \|^2, \quad k = 0, 1, \ldots$$

By virtue of the condition (b), the sequence $\{M\varphi(\mathbf{x}_k)\}$ is bounded from below, and, since the above inequality holds for all the indices k, we have $M \| \varphi'(\mathbf{x}_k) \|^2 \to 0$ for $k \to \infty$, and hence the condition (c) implies that

$$\lim_{k \to \infty} P\{\rho(\mathbf{x}_k, X^*) \geqslant \varepsilon\} = 0 \text{ for any } \varepsilon > 0. \quad \triangle$$

10.6.6. Corollary. *If the function $\varphi(\mathbf{x})$ satisfies the conditions of the foregoing theorem and, besides, is convex, then*

$$\rho(\mathbf{x}_k, X^*) \xrightarrow[\text{a.c.}]{} 0, \quad k \to \infty.$$

Indeed, as was proved, the sequence $\{\rho(x_k, X^*)\}$ converges in probability to zero, and, consequently, so does the sequence $\{\mu_k\}$, where $\mu_k = \varphi(x_k) - \varphi(x^*)$. Further, there exists a subsequence $\{\mu_{k_j}\}$ converging to zero almost certainly, and therefore, since the sequence $\{\mu_k\}$ is monotonically non-decreasing, we have $\mu_k \underset{\text{a.c.}}{\to} 0$, $k \to \infty$, and, consequently, in view of the condition (c) and Theorem 9.2.5, $\rho(x_k, X^*) \underset{\text{a.c.}}{\to} 0$, $k \to \infty$, where X^* is the set of points of minima of the function $\varphi(x)$. $\triangle$

10.6.7. To estimate the rate of convergence of the minimization process we need the following

Lemma. *The sequence of random variables* $\{\alpha_k\}$ *satisfies the inequality*

$$P\left\{\sum_{j=0}^{k-1} \alpha_j^2 \geqslant \frac{1}{2}\alpha^2 pk\right\} \geqslant 1 - \frac{4}{\alpha^4 p^2 k}, \quad k = 1, 2, \ldots$$

Proof. Consider an auxiliary sequence of random variables $\{\xi_k\}$ such that $P\{\xi_k = \alpha^2\} = p$ and $P\{\xi_k = 0\} = 1 - p$. Here α and p are the constants involved in (10.43). We shall prove that

$$P\left\{\sum_{j=0}^{k-1} \alpha_j^2 \geqslant Ck\right\} \geqslant P\left\{\sum_{j=0}^{k-1} \xi_j \geqslant Ck\right\} \tag{10.44}$$

for all indices $k = 1, 2, \ldots$ and any $C \in (-\infty, +\infty)$.

Let $k = 1$. From (10.43) we obtain

$$P\{\alpha_0^2 \geqslant C\} \geqslant \begin{cases} 0 & \text{for } C > \alpha^2, \\ p & \text{for } 0 < C \leqslant \alpha^2, \\ 1 & \text{for } C \leqslant 0, \end{cases}$$

and, consequently, $P\{\alpha_0^2 \geqslant C\} \geqslant P\{\xi_0 \geqslant C\}$.

We now state the induction hypothesis that (10.44) holds for all $k \leqslant m$.

Let $k = m + 1$. Denote $S_k \equiv \sum_{j=0}^{k-1} \alpha_j^2$. By the theorem of total probability,

$$P\{S_{m+1} \geqslant C(m+1)\} = P\{\alpha_m^2 + S_m \geqslant C(m+1)\}$$
$$= P\{S_{m+1} \geqslant C(m+1) \mid S_m \geqslant C(m+1)\} P\{S_m \geqslant C(m+1)\}$$
$$+ P\{S_{m+1} \geqslant C(m+1) \mid C(m+1) - \alpha^2 \leqslant S_m < C(m+1)\}$$

$$\times P\{C(m+1)-\alpha^2 \leqslant S_m < C(m+1)\}$$
$$+ P\{S_{m+1} \geqslant C(m+1) \mid S_m < C(m+1)-\alpha^2\}$$
$$\times P\{S_m < C(m+1)-\alpha^2\}.$$

Since the random variables α_k^2 are nonnegative, we have

$$P\{S_{m+1} \geqslant C(m+1) \mid S_m \geqslant C(m+1)\} = 1.$$

Moreover, the condition (10.43) implies

$$P\{S_{m+1} \geqslant C(m+1) \mid C(m+1)-\alpha^2 \leqslant S_m < C(m+1)\}$$
$$\geqslant P\{\alpha_m^2 \geqslant \alpha^2 \mid C(m+1)-\alpha^2 \leqslant S_m < C(m+1)\} \geqslant p.$$

Therefore

$$P\{S_{m+1} \geqslant C(m+1)\}\, P\{S_m \geqslant C(m+1)\}$$
$$+ pP\{C(m+1)-\alpha^2 \leqslant S_m < C(m+1)\}.$$

From the theorem on addition of probabilities it follows that

$$P\{C(m+1)-\alpha^2 \leqslant S_m < C(m+1)\}$$
$$= P\{S_m \geqslant C(m+1)-\alpha^2\} - P\{S_m \geqslant C(m+1)\},$$

and consequently

$$P\{S_{m+1} \geqslant C(m+1)\}$$
$$\leqslant (1-p)\, P\{S_m \geqslant C(m+1)\} + pP\{S_m \geqslant C(m+1)-\alpha^2\}$$
$$= (1-p)\, P\{S_m \geqslant C_1 m\} + pP\{S_m \geqslant C_2 m\}, \quad (10.45)$$

where $C_1 = C\frac{m+1}{m}$, $C_2 = +\frac{C-\alpha^2}{m}$.

For the independent random variables ξ_k we put $R_k = \sum_{j=0}^{k-1} \xi_j$. Then

$$P\{R_{m+1} \geqslant C(m+1)\}$$
$$= P\{\xi_m = 0,\ R_m \geqslant C(m+1)\} + P\{\xi_m = \alpha^2,\ R_m$$
$$\geqslant C(m+1)-\alpha^2\} = (1-p)\, P\{R_m \geqslant C(m+1)\}$$
$$+ pP\{R_m \geqslant C(m+1)-\alpha^2\} \leqslant (1-p)\, P\{R_m \geqslant C_1 m\}$$
$$+ pP\{R_m \geqslant C_2 m\}. \quad (10.46)$$

By the induction hypothesis, (10.44) holds for $k = m$ and any $C \in (-\infty, +\infty)$ and, in particular, for $C = C_1$ and

$C = C_2$. Therefore from (10.45) we obtain

$$P\{S_{m+1} \geqslant C(m+1)\} \geqslant (1-p)\,P\{R_m \geqslant C_1 m\} + pP\{R_m \geqslant C_2 m\}.$$

This and (10.46) imply (10.44).

For R_k we have an obvious relation

$$P\left\{R_k \leqslant \frac{1}{2}\alpha^2 pk\right\} = P\left\{\sum_{j=0}^{k-1}(\xi_j - \alpha^2 p) \leqslant -\frac{1}{2}\alpha^2 pk\right\}.$$

By the Kolmogorov inequality, the right-hand probability does not exceed $4/(\alpha^4 p^2 k)$. It remains to put $C = \alpha^2 p/2$ in (10.44) to prove the assertion of the lemma. △

10.6.8. Estimation of the rate of convergence.

Theorem. *If the function* $\varphi(\mathbf{x})$ *is convex and satisfies the conditions*

(a) $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$,

(b) $X^* \neq \varnothing$

(c) *for any* $\varepsilon > 0$ *there is* $\delta = \delta(\varepsilon) > 0$ *such that* $\|\varphi'(\mathbf{x})\| \geqslant \delta$ *for all* $\mathbf{x}$ *such that* $\rho(\mathbf{x}, X^*) \geqslant \varepsilon$, *then*

$$P\{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}^*) \geqslant C_1/k\} \geqslant 1 - C_2/k, \quad k = 1, 2, \ldots, \tag{10.47}$$

where $C_1 > 0$ *and* $C_2 > 0$ *are some positive constants.*

Proof. Theorem 9.3.4 implies the inequality

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}^*) \leqslant \eta^2\left(\sum_{j=0}^{k-1}\frac{\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1})}{\|\varphi'(\mathbf{x}_j)\|^2}\right)^{-1}, \quad k = 1, 2, \ldots,$$

and, by Lemma 9.4.8, we have

$$\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1}) \geqslant h\alpha_j^2 \|\varphi'(\mathbf{x}_j)\|^2, \quad h = \lambda/(2L).$$

Therefore

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}^*) \leqslant \frac{\eta^2}{h}\left(\sum_{j=0}^{k-1}\alpha_j^2\right)^{-1}, \quad k = 1, 2, \ldots$$

Lemma 10.6.7 implies the inequality

$$P\left\{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}^*) \leqslant \frac{\eta^2}{h}\left(\sum_{j=0}^{k-1}\alpha_j^2\right)^{-1} \leqslant \frac{2\eta^2}{h\alpha^2 pk}\right\} \geqslant 1 - \frac{4}{\alpha^4 p^2 k}, \quad (10.48)$$

whence it follows that the theorem is true for $C_1 = 2\eta^2/(h\alpha^2 p)$ and $C_2 = 4/(\alpha^4 p^2)$. $\triangle$

10.6.9. Corollary. *If the function* $\varphi(\mathbf{x})$ *is strongly convex and belongs to the class* $C^{1,1}(E_n)$, *then*

$$P\left\{\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leqslant \frac{2}{\rho}\mu_0 \exp\{-C_3 k\}\right\} \geqslant 1 - \frac{C_2}{k}, \quad C_3 > 0, \quad (10.49)$$

where ρ *is the parameter of strong convexity.*

Indeed, using Theorem 9.3.5 and the foregoing estimate for the difference $\varphi(\mathbf{x}_j) - \varphi(\mathbf{x}_{j+1})$ we arrive at the inequality

$$\|\mathbf{x}_k - \mathbf{x}^*\|^2 \leqslant \frac{2}{\rho}\mu_0 \exp\left\{-\rho h \sum_{j=0}^{k-1}\alpha_j^2\right\}, \quad k = 1, 2, \ldots$$

Finally, applying Lemma 10.6.7 we obtain (10.48) for $C_3 = \rho h \alpha p/2$. $\triangle$

10.6.10. Examples of random vectors $\mathbf{s}_k$ $(k = 0, 1, \ldots)$.

1. *Random coordinate-wise descent.* Let $\mathbf{s}_k = \mathbf{e}_{j(k)}$ where $\mathbf{e}_{j(k)}$ is the jth coordinate vector. The index $j = j(k)$ is chosen randomly from the n numbers $\{1, 2, \ldots, n\}$ so that $P\{j(k) = i\} = 1/n$ $(i = \overline{1, n})$. Then, as can easily be seen,

$$P\left\{\alpha_k^2 \geqslant \frac{1}{n}\right\} = P\left\{\frac{\langle\varphi'(\mathbf{x}_k), \mathbf{s}_k\rangle^2}{\|\varphi'(\mathbf{x}_k)\|^2\|\mathbf{s}_k\|^2} \geqslant \frac{1}{n}\right\}.$$

Taking into account that $\langle\varphi'(\mathbf{x}_k), \mathbf{s}_k\rangle^2 = (\varphi'_{j(k)}(\mathbf{x}_k))^2$, $\|\mathbf{s}_k\| = 1$, we can write

$$P\left\{\alpha_k^2 \geqslant \frac{1}{n}\right\} = P\left\{(\varphi'_{j(k)}(\mathbf{x}_k))^2\left(\sum_{i=1}^{n}(\varphi'_i(\mathbf{x}_k))^2\right)^{-1} \geqslant \frac{1}{n}\right\}$$

$$\geqslant P\left\{(\varphi'_{j(k)}(\mathbf{x}_k))^2 \geqslant \frac{\|\varphi'(\mathbf{x}_k)\|^2}{n}\right\}$$

$$\geqslant P\left\{(\varphi'_{j(k)}(\mathbf{x}_k))^2 = \max_{i=\overline{1,n}}(\varphi'_i(\mathbf{x}_k))^2\right\} = \frac{1}{n}.$$

In practice the method of random coordinate-wise descent is applied in combination with certain heuristic techniques to accelerate the minimization process. In this way, when selecting the direction of descent, one excludes from the consideration "non-promising" directions; for instance, when selecting the index $j(k)$, one excludes the direction $-s_k = -s_{k-1}$ which has already been used at the foregoing step. Thus, the index $j(k)$ is chosen from the $(n-1)$ numbers $\{1, 2, \ldots, j(k-1)-1, j(k-1)+1, \ldots, n\}$.

As a result of a series of trials, some directions can be rejected, at least temporarily, as non-promising, and the index $j(k)$ is chosen from a smaller set of numbers.

Finally, the method of random coordinate-wise descent is often used together with other approaches by forming the so-called hybrid methods. For example, after a number $p \leqslant n$ is chosen for some reasons based on the information gained during the calculation process and the points $\mathbf{x}_0, \ldots, \mathbf{x}_p$ are constructed, the direction $\mathbf{x}_p - \mathbf{x}_0$ is taken as the direction of descent $-s_{p+1}$, and then the procedure of random coordinate-wise descent is repeated.

2. *The random descent method.* Let s_k be a random point having a uniform probability distribution on the sphere $S = \{s \in E_n: \|s\| = 1\}$.

Definition. A random point $s \in E_n$ is said to be *uniformly distributed* over a squarable surface S if $P\{s \in S'\} = \text{mes } S'/\text{mes } S$ $(S' \subseteq S)$ for any measurable subset S'. Here the symbol mes is understood as area.

The definition implies that the probability density of the vector s is constant and is equal to $1/\text{mes } S$.

Since s_k are independent random vectors uniformly distributed over the sphere S and hence possessing a constant probability density, for some $p > 0$ depending on the dimension n of the space E_n we have

$$P\left\{\left\| s_k - \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\| \leqslant \frac{1}{2}\right\} \geqslant p > 0.$$

However, if

$$\left\| s_k - \frac{\varphi'(s_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\| \leqslant \frac{1}{2},$$

then

$$\alpha_k^2 = \frac{\langle \varphi'(\mathbf{x}_k), s_k\rangle^2}{\|\varphi'(\mathbf{x}_k)\|^2} = \left(1 - \left\langle \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}, -s_k + \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\rangle\right)^2$$
$$\geqslant \left(1 - \left\| s_k - \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\|\right)^2 \geqslant \frac{1}{4}.$$

Thus,

$$P\left\{\left(1 - \left\| s_k - \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\|\right)^2 \geqslant \frac{1}{4}\right\} \geqslant p > 0,$$

and, at the same time,

$$P\left\{\alpha_k^2 \geqslant \left(1 - \left\| s_k - \frac{\varphi'(\mathbf{x}_k)}{\|\varphi'(\mathbf{x}_k)\|}\right\|\right)^2\right\} = 1,$$

whence $P\left\{\alpha_k^2 \geqslant \frac{1}{4} \,\middle|\, \mathbf{x}_k, \ldots, \mathbf{x}_0\right\} \geqslant p > 0.$

We see that the condition (10.43) holds for $\alpha = 1/2$ and some $p > 0$.

3. As an example of application of dependent random vectors, we consider the so-called *method of random search with learning*. The direction s_k will be determined in each iteration by the formula

$$s_k = \frac{1}{\|v_k + w_k\|}(v_k + w_k), \quad \|w_k\| \leqslant c < 1, \quad k = 0, 1, \ldots,$$

where $\{v_k\}$ is a sequence of independent random vectors uniformly distributed over the unit sphere with center at the origin, w_k is the instructing vector whose choice is carried out taking into account the information about the minimized function and the set X gained in the foregoing iterations.

We shall prove that for any fixed $\bar{s}$ ($\|\bar{s}\| = 1$) and any $\alpha \in (0, 1)$ there holds $P\{\langle s_k, \bar{s}\rangle \geqslant \alpha \mid x_k, \ldots, x_0\} \geqslant p > 0$, $k = 0, 1, \ldots$, i.e. the conditions (10.43) are fulfilled for $\bar{s} = \varphi'(x_k)/\|\varphi'(x_k)\|$.

In what follows, to avoid cumbersome expressions, the index k is omitted. Consider the vector $y = -w + (\delta + \langle \bar{s}, w\rangle)\bar{s}$ where $\delta = \sqrt{1 + \langle \bar{s}, w\rangle^2 - \|w\|^2}$. It can readily be seen that $\|y\| = 1$ and $\delta + \langle \bar{s}, w\rangle \geqslant \sqrt{1 - c^2 + \langle \bar{s}, w\rangle^2} + \langle \bar{s}, w\rangle > 0$. Therefore, $\langle y + w, \bar{s}\rangle = \delta + \langle \bar{s}, w\rangle = \|y + w\|$. We take $\varepsilon = (1 - \alpha)^2 \times (1 - c)^2/8$ and assume that the event $\langle v, y\rangle \geqslant 1 - \varepsilon$ has occurred. It should be noted that

$$\|v - y\|^2 = 2(1 - \langle v, y\rangle) \leqslant 2\varepsilon,$$

$$\langle s, \bar{s}\rangle = \frac{1}{\|v + w\|}(\langle y + w, \bar{s}\rangle + \langle v - y, \bar{s}\rangle)$$

$$\geqslant \frac{1}{\|v + w\|}(\|y + w\| - \|v - y\|) \geqslant \frac{1}{\|v + w\|}(\|y + w\| - \sqrt{2\varepsilon}) > 0.$$

Since $\|v + w\| \leqslant \|y + w\| + \|v - y\| \leqslant \|y + w\| + \sqrt{2\varepsilon}$, we have

$$\langle s, \bar{s}\rangle \geqslant \frac{\|y + w\| - \sqrt{2\varepsilon}}{\|y + w\| + \sqrt{2\varepsilon}} = 1 - \frac{2\sqrt{2\varepsilon}}{\|y + w\| + \sqrt{2\varepsilon}} > 1 - \frac{2\sqrt{2\varepsilon}}{\|y\| - \|w\|}$$

$$\geqslant 1 - \frac{2\sqrt{2\varepsilon}}{1 - c} = \alpha.$$

Since $\{v_k\}$ is a sequence of independent uniformly distributed random vectors of unit length, for any unit vector a ($\|a\| = 1$) and any $\varepsilon \in (0, 1)$ there is a number $p = p(\varepsilon) \in (0, 1)$ depending neither on the index k nor on the vector a and such that $P\{\langle v_k, a\rangle \geqslant 1 - \varepsilon\} = p$. Therefore

$$P\{\langle s_k, \bar{s}\rangle \geqslant \alpha \mid x_k, \ldots, x_0\} \geqslant P\{\langle v_k, y_k\rangle \geqslant 1 - \varepsilon\} = p, \quad k = 0, 1, \ldots$$

10.6.11. The random descent method for finding stationary points in the general nonlinear programming

problem. For the problem of minimizing a function $\varphi(x)$ on a set $X = \{x: f_i(x) \geqslant 0, i = \overline{1, m}\}$ the random descent method is a process of constructing a sequence $\{x_k\}$ whose elements satisfy the following relations $x_{k+1} = x_k - \beta_k s_k$, $\beta_k \geqslant 0$, $[x_{k+1}, x_k] \subset X$, and $\varphi(x_{k+1}) \leqslant \varphi(x_k)$. Here s_k is a random vector of unit length. Various versions of the method differ in the ways the direction s_k and the step β_k are selected.

From the computational standpoint the method is quite elementary since the direction s_k is found in a simple way using a generator of random numbers. However, in contrast to the corresponding method in unconditional minimization problems, here some effort is required to verify whether the direction $-s_k$ is feasible at the point x_k. To this end, it is found out whether the point $x_k - \beta s_k$ belongs to the set X for a sufficiently small β, and only after that, if $-s_k$ is a feasible direction, it is checked whether the function $\varphi(x)$ decreases along this direction or not. Only after the positive answer to the question is obtained is the descent procedure realized.

As in Sec. 10.4.3, we denote by ζ_k the distance from the point x_k to the nearest boundary point of the set X in the direction $-s_k$.

10.6.12. The scheme of the method. As an initial approximation an arbitrary point x_0 belonging to the set X can be taken. Let the kth iteration result in the calculation of $x_k \in X$. We shall describe the $(k+1)$th iteration.

(1) A random direction s_k is found and it is verified whether $-s_k$ is a feasible direction at the point x_k.

(2) If the direction $-s_k$ is not feasible, one puts $x_{k+1} = x_k$ and passes to (1) for $k = k + 1$.

(3) If $-s_k$ is a feasible direction, then as β_k one takes any number β belonging to the segment $[0, \zeta_k]$ and satisfying the condition

$$\varphi(x_k - \beta_k s_k) \leqslant (1 - \lambda_k)\varphi(x_k) + \lambda_k \omega_k,$$
$$0 < \lambda \leqslant \lambda_k \leqslant 1, \quad \omega_k = \inf_{0 \leqslant \beta \leqslant \zeta_k} \varphi(x_k - \beta s_k). \quad (10.26)$$

Finally, one puts $x_{k+1} = x_k - \beta_k s_k$ and passes to (1) for $k = k + 1$.

Remark. The procedure in (3) also includes the situation when $\varphi(x_k) \leqslant \varphi(x_k - \beta s_k)$ for all $\beta \in [0, \zeta_k]$. In this case $\beta_k = 0$ and $x_{k+1} = x_k$.

10.6.13. The set of stationary points. We consider an auxiliary problem (see Sec. 10.4.2) of finding a number σ and a vector s such that

$$\begin{gathered} \sigma \to \max, \\ \langle f_i'(x), s\rangle + \sigma \leqslant 0, \quad i \in I, \\ -\langle \varphi'(x), s\rangle + \sigma \leqslant 0, \\ \langle s, s\rangle \leqslant 1. \end{gathered} \tag{10.25}$$

As in Sec. 10.4.2, we introduce the following notation. Let $\varepsilon > 0$, then

$$I(x, \varepsilon) = \{i: 0 \leqslant f_i(x) \leqslant \varepsilon\}, \quad I(x, 0) = \{i: f_i(x) = 0\};$$

$\tilde{\sigma}(x, \varepsilon)$, $\tilde{s}(x, \varepsilon)$ is the solution to the problem (10.25) for $I = I(x, \varepsilon)$; $\tilde{\sigma}(x, 0)$, $\tilde{s}(x, 0)$ is the solution to the problem (10.25) for $I = I(x, 0)$; $X_0 = \{x \in X: \varphi(x) \leqslant \varphi(x_0)\}$; and $X^* = \{x^* \in X_0; \tilde{\sigma}(x^*, 0) = 0\}$. In accordance with Sec. 10.4.7, X^* is the set of stationary points where the necessary conditions for local minimum are fulfilled.

Assumptions. The convergence of the method will be proved under the following assumptions:

(1) s_k are independent random vectors uniformly distributed over the sphere $S = \{s: \|s\| = 1\}$,

(2) $\varphi(x)$ and $f_i(x)$ $(i = \overline{1, m})$ are functions of class $C^{1,1}(X)$.

(3) the set X_0 is bounded.

It should be noted that the investigation of the convergence will be essentially based on the material of Sec. 10.4. In the case under consideration the assumption that the set X_0 is bounded implies that the conditions (2)-(4) in Sec. 10.4.6 hold: $\|f_i'(x)\| \leqslant M$ for all $x \in X_0$, $X^* \neq \varnothing$, and $\inf\limits_{x \in X} \varphi(x) > -\infty$.

We shall denote by $\bar{\sigma}$, $\bar{s}$ the solution to the problem (10.25) for $x = \bar{x}$ and $I = I(\bar{x}, 0)$.

10.6.14. Lemma. *If* $\bar{\sigma} > 0$, *then there are* $\bar{\varepsilon} > 0$, $\delta > 0$, *and* $\alpha \in (0, 1)$ *such that for any* $\varepsilon \in [0, \bar{\varepsilon}]$, $x \in U_\delta(\bar{x}) = \{x \in X: \| x - \bar{x} \| \leqslant \delta\}$ *and any direction* s ($\| s \| = 1$) *satisfying the condition* $\langle s, \bar{s} \rangle \geqslant 1 - \alpha$ *the constraints in the problem* (10.25) *are fulfilled for* $\sigma = \bar{\sigma}/4$ *and* $I = I(x, \varepsilon)$.

Proof. Let $\bar{\varepsilon}$ and δ be such that the inequalities $\| \varphi'(x) - \varphi'(\bar{x}) \| \leqslant \bar{\sigma}/2$, $\| f_i'(x) - f_i'(\bar{x}) \| \leqslant \bar{\sigma}/2$ ($i = \overline{1, m}$) hold. Then for any $x \in U_\delta(\bar{x})$ and any $i \in I(x, \varepsilon) \subset I(\bar{x}, 0)$ we have

$$0 \geqslant \langle f_i'(\bar{x}), \bar{s} \rangle + \bar{\sigma} = \langle f_i'(x), \bar{s} \rangle + \bar{\sigma} + \langle f_i'(\bar{x}) - f_i'(x), \bar{s} \rangle \geqslant \langle f_i'(x), \bar{s} \rangle + \bar{\sigma} - \| f_i'(\bar{x}) - f_i'(x) \| \geqslant \langle f_i'(x), \bar{s} \rangle + \bar{\sigma}/2 \tag{10.50}$$

and, similarly

$$0 \geqslant -\langle \varphi'(x), \bar{s} \rangle + \bar{\sigma}/2. \tag{10.51}$$

Let

$$\alpha = \bar{\sigma}^2/(32M_1^2), \qquad M_1 = \max\left\{\max_{x \in X_0} \| \varphi'(x) \|, M\right\}.$$

We note that $\bar{\sigma}/2 \leqslant \langle \varphi'(x), \bar{s} \rangle \leqslant \| \varphi'(x) \| \leqslant M_1$. Further, $\langle f_i'(x), s \rangle = \langle f_i'(x), \bar{s} \rangle + \langle f_i'(x), s - \bar{s} \rangle$. However,

$$\langle f_i'(x), s - \bar{s} \rangle \leqslant M_1 \| s - \bar{s} \| \leqslant M_1 \sqrt{\| s \|^2 - 2\langle s, \bar{s} \rangle + \| \bar{s} \|^2} \leqslant M_1 \sqrt{2 - 2(1 - \alpha)} = \bar{\sigma}/4,$$

whence, taking into account (10.50) and the foregoing inequality we obtain $\langle f_i'(x), s \rangle \leqslant -\bar{\sigma}/2 + \bar{\sigma}/4 = -\bar{\sigma}/4$ and, similarly, $\langle \varphi'(x), s \rangle \leqslant \langle \varphi'(x), \bar{s} \rangle + \langle \varphi'(x), s - \bar{s} \rangle \leqslant \bar{\sigma}/2 - M_1 \| s - \bar{s} \| \geqslant \bar{\sigma}/4$. $\triangle$

10.6.15. Lemma. *For any* $\gamma \geqslant 0$ *there are* $\sigma_\gamma > 0$ *and* $\zeta_\gamma > 0$ *such that for all* $\bar{x} \in X_0 \setminus U_\gamma(X^*)$ *and* $\beta \in [0, \zeta_\gamma]$ *we have* $\bar{\sigma} > \sigma_\gamma > 0$ *and* $x - \beta\bar{s} \in X$. *Here* $U_\gamma(X^*) = \{x \in X_0 \mid \rho(x, X^*) \leqslant \gamma\}$.

Proof. We first prove that $\sigma_\gamma > 0$ exists for any $\gamma > 0$. Assume the contrary: let $\bar{\sigma}_k \to 0$, $k \to \infty$, for a sequence $\{y_k\}$ where the pair $\bar{\sigma}_k$, $\bar{s}_k$ is the solution to the problem

(10.25) for $\mathbf{x} = \mathbf{y}_k$, $I = I(\mathbf{y}_k, 0)$ and $\mathbf{y}_k \in X_0 \setminus U_\gamma(X^*)$ for some $\gamma > 0$.

The set X_0 being bounded, we can assume, without loss of generality, that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{y}$, where $\rho(\mathbf{y}, X^*) \geqslant \gamma$. Let $\tilde{\sigma}$, $\tilde{\mathbf{s}}$ be the solution to (10.25) for $\mathbf{x} = \mathbf{y}$ and $I = I(\mathbf{y}, 0)$. Since $\mathbf{y} \notin X^*$, we have $\tilde{\sigma} > 0$. By virtue of Lemma 10.6.14, there are $\tilde{\varepsilon} > 0$ and $\tilde{\delta} > 0$ such that for all $\mathbf{x} \in U_{\tilde{\delta}}(\mathbf{y})$ the pair $\sigma = \tilde{\sigma}/4$, $\mathbf{s} = \tilde{\mathbf{s}}$ satisfies the constraints in the problem (10.25) for $I = I(\mathbf{x}, \tilde{\varepsilon})$. Since $\mathbf{y}_k \to \mathbf{y}$ for $k \to \infty$, the vectors $\mathbf{y}_k$ belong to $U_{\tilde{\delta}}(\mathbf{y})$ for all sufficiently large k, and hence, by virtue of the same lemma, in the pair $\bar{\sigma}_k$, $\bar{\mathbf{s}}_k$ there must be $\bar{\sigma}_k \geqslant \tilde{\sigma}/4 > 0$ since σ_k is a component of the solution to (10.25) and therefore has the maximum value. The contradiction we have obtained proves the existence of $\sigma_\gamma > 0$ for any $\gamma > 0$.

We now prove the existence of $\zeta_\gamma > 0$. Assume the contrary: let $\bar{\zeta}_k \to 0$, $k \to \infty$, for some sequence $\{\mathbf{y}_k\}$, $\mathbf{y}_k \in X_0 \setminus U_\gamma(X^*)$. Here $\gamma > 0$, $\bar{\zeta}_k$ is the distance from the point $\mathbf{y}_k$ to the boundary of the set X along the direction $-\bar{\mathbf{s}}_k$, and the pair $\bar{\sigma}_k$, $\bar{\mathbf{s}}_k$ is the solution to the problem (10.25) for $\mathbf{x} = \mathbf{y}_k$, $I = I(\mathbf{y}_k, 0)$. As above, we shall assume that $\lim_{k\to\infty} \mathbf{y}_k = \mathbf{y}$ and $\rho(\mathbf{y}, X^*) \geqslant \gamma$. By virtue of Lemma 10.4.10, there are $\bar{\varepsilon} > 0$ and $\delta > 0$ such that for all $\mathbf{y}_k \in U_\delta(\mathbf{y})$ the inclusion $I(\mathbf{y}_k, \bar{\varepsilon}) \subset I(\mathbf{y}, 0)$ takes place. Then Lemma 10.6.14 implies that the constraints in the problem (10.25) are fulfilled for $\sigma = \frac{1}{4}\tilde{\sigma}$, $\mathbf{s} = \tilde{\mathbf{s}}$, where the pair $\tilde{\sigma}$, $\tilde{\mathbf{s}}$ is the solution to (10.25) for $\mathbf{x} = \mathbf{y}$, $I = I(\mathbf{y}_k, \bar{\varepsilon})$. By Lemma 10.4.8, the relations (10.30) and (10.31) imply $\bar{\zeta}_k \geqslant \min\{\bar{\varepsilon}/\mu, \sigma_\gamma/L\} > 0$. We have arrived at a contradiction proving the lemma. △

10.6.16. Theorem. *The sequence of random points* $\{\mathbf{x}_k\}$ *converges in probability to the set* X^*.

Proof. We fix an arbitrary number $\gamma > 0$ and prove that $P\{\rho(\mathbf{x}_k, X^*) \geqslant \gamma\} \to 0$, $k \to \infty$. Assume that $\mathbf{x}_k \in X_0 \setminus U_\gamma(X^*)$. Then by Lemma 10.6.15, there are $\sigma_\gamma > 0$

and $\zeta_\nu > 0$ such that $\bar{\sigma}_k \geqslant \sigma_\nu$ and $\bar{\zeta}_k \geqslant \zeta_\nu$. Here $\bar{\sigma}_k$, $\bar{\mathbf{s}}_k$ is the solution to (10.25) for $\mathbf{x} = \mathbf{x}_k$ and $I = I(\mathbf{x}_k, 0)$. Since $\bar{\sigma}_k > 0$, by Lemma 10.6.14, there is $\bar{\varepsilon}_k > 0$ such that any vector $\mathbf{s}_k$ satisfying the condition $\langle \mathbf{s}_k, \bar{\mathbf{s}}_k \rangle \geqslant 1 - \alpha$ satisfies the constraints in the problem (10.25) for $\sigma = \frac{1}{4}\bar{\sigma}_k$ and $I = I(\mathbf{x}_k, \bar{\varepsilon}_k)$. By virtue of the choice of the parameter β_k (see the conditions (10.26)), for all $\beta \in [0, \bar{\zeta}_k]$ we have $\mathbf{x}_k - \beta \mathbf{s}_k \in X_0$ and (see Sec. 10.4.9)

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant \lambda(\varphi(\mathbf{x}_k) - \omega_k) \geqslant \lambda(\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta \mathbf{s}_k))$$

$$\geqslant \lambda\left(\langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle \beta - \frac{1}{2} L\beta^2 \| \mathbf{s}_k \|^2\right) \geqslant \lambda\left(\beta\sigma - \frac{1}{2}\beta^2 L\right).$$

Taking $\beta = \min\{\bar{\zeta}_k, \sigma/L\}$ we obtain, taking into consideration that $\bar{\zeta}_k > \zeta_\nu$ and $\sigma \geqslant \frac{1}{4}\sigma_\nu$, the relation

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta_k \mathbf{s}_k) \geqslant \varkappa \tag{10.52}$$

holding for all $\mathbf{s}_k$ such that $\langle \mathbf{s}, \bar{\mathbf{s}}_k \rangle \geqslant 1 - \alpha$, where $\varkappa = \min\left\{\zeta_\nu, \frac{1}{8}(\sigma_\nu/L)\right\}$.

Let ξ_k be random variables such that

$$\xi_k = \begin{cases} 1 & \text{if } \mathbf{x}_k \in X_0 \setminus U_\delta(X^*), \\ 0 & \text{if } \mathbf{x}_k \in U_\delta(X^*). \end{cases}$$

Since the points $\mathbf{s}_k$ are uniformly distributed over the sphere $S = \{\mathbf{s}: \| \mathbf{s} \| = 1\}$, it follows from (10.52) that

$$P\{\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \geqslant \varkappa \mid \xi_k = 1\} \geqslant P\{\langle \mathbf{s}_k, \bar{\mathbf{s}}_k \rangle \geqslant 1 - \alpha\}$$

$$\geqslant P\{\| \mathbf{s}_k - \bar{\mathbf{s}}_k \|^2 \leqslant 2\alpha\} \geqslant p > 0,$$

whence $M(\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) \mid \mathbf{x}_k, \ldots, \mathbf{x}_0) \geqslant \varkappa p \xi_k$. By analogy with Sec. 10.6.5, using this inequality we readily derive the condition

$$M\varphi(\mathbf{x}_{k+1}) \leqslant \varphi(\mathbf{x}_0) - \varkappa p \sum_{j=0}^{k} M\xi_j.$$

Since, by the hypothesis, the set X_0 is bounded, the function $\varphi(\mathbf{x})$ on X_0 is also bounded, and, consequently, $M\xi_k \to 0$, $k \to \infty$, and hence $P\{\rho(\mathbf{x}_k, X^*) \geqslant \gamma\} \to 0$, $k \to \infty$. $\triangle$

10.6.17. On the convergence of the method in the case of convex programming problems. In unconditional minimization problems the convergence in probability of the sequence $\{x_k\}$ to the set of stationary points implied, in the convex case, the (almost certain) convergence of the sequence $\{\varphi(x_k)\}$ to $\varphi^* = \min \varphi(x)$ and, consequently, the (almost certain) convergence of $\{x_k\}$ to the solution set. In problems with constraints the condition that the function $\varphi(x)$ and the set X are convex does not imply an analogous assertion. To prove that $\varphi(x_k) \xrightarrow[\text{a.c.}]{} \varphi^*$, $k \to \infty$, and $\rho(x_k, X^*) \xrightarrow[\text{a.c.}]{} 0$, $k \to \infty$, it is necessary to impose an additional requirement that the set X be regular in Slater's sense.

10.6.18. Assumptions:

(1) s_k are independent random vectors uniformly distributed over the sphere $S = \{s: \|s\| = 1\}$,

(2) the set X is convex, closed, and regular in Slater's sense, i.e. there is a ball R of radius r, with center at a point w, entirely belonging to the set X: $R = \{x: \rho(x, w) \leqslant r\} \subset X$.

(3) $\varphi(x)$ is a convex function of class $C^{1,1}(X)$,

(4) the set X_0 is bounded.

Remark. It can easily be shown that the condition of regularity (2) is equivalent to that introduced in Secs. 3.4.2 and 3.4.3.

To prove the convergence of the procedure of random descent it is necessary to consider a number of properties.

10.6.19. Property A. *If $\bar{x} \in X$ is not an optimal point, then there is a vector $\bar{y} \in X$ satisfying the inequality*

$$\langle \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle = \eta > 0.$$

Proof. Let $\bar{y}$ be the projection of the point $z = \bar{x} - \varphi'(\bar{x})$ on the set X. By virtue of the properties of projections, we have $\langle \bar{x} - z, \bar{x} - \bar{y} \rangle \geqslant \|\bar{x} - \bar{y}\|^2$, i.e. $\langle \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle \geqslant \|\bar{x} - \bar{y}\|^2$. This implies Property A since the relation $\bar{x} = \bar{y}$ is a necessary and sufficient condition for $\bar{x}$ to be optimal (Sec. 10.1.2). $\triangle$

Below $\bar{x}$ and $\bar{y}$ everywhere denote points possessing Property A.

10.6.20. Property B. *There is* $\delta > 0$ *such that the sets*

$$U_\delta(\bar{x}) = \{x \in X: \| x - \bar{x} \| \leqslant \delta\},$$
$$U_\delta(\bar{y}) = \{y \in X: \| y - \bar{y} \| \leqslant \delta\}$$

do not intersect and there exist n-dimensional spheres R_1 *and* R_2 *of radii* r_1 *and* r_2, *respectively, such that* $R_1 \subset U_\delta(\bar{x})$ *and* $R_2 \subset U_\delta(\bar{y})$.

The proof follows from Property A (since $\bar{x} \neq \bar{y}$) and the assumption (2)*. $\triangle$

10.6.21. Property C. *For* $\eta = \langle \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle$ *there is* $\delta > 0$ *such that for all* $x \in U_\delta(\bar{x})$ *and* $y \in U_\delta(\bar{y})$ *the inequality*

$$\langle \varphi'(x), x - y \rangle \geqslant \frac{1}{2}\eta$$

is fulfilled.

Proof. Denote $\gamma = \max_{x \in U_\delta(\bar{x})} \| \varphi'(x) \|$. The assumption (3) guarantees that $\gamma < +\infty$. We take

$$\delta \leqslant \frac{\eta}{2(L \| \bar{x} - \bar{y} \| + 2\gamma)}.$$

Let $u = x - \bar{x}$ and $v = y - \bar{y}$. Then

$$\langle \varphi'(x), x - y \rangle = \langle \varphi'(\bar{x} + u), \bar{x} - \bar{y} \rangle + \langle \varphi'(\bar{x} + u), u - v \rangle$$
$$\geqslant \langle \varphi'(\bar{x} + u), \bar{x} - \bar{y} \rangle - | \langle \varphi'(\bar{x} + u), u - v \rangle |.$$

Further,

$$| \langle \varphi'(\bar{x} + u), u - v \rangle | \leqslant \| \varphi'(\bar{x} + u) \| \| u - v \|$$
$$\leqslant \gamma (\| u \| + \| v \|) \leqslant 2\gamma\delta.$$

However, in view of the assumption (3) and Property A, we obtain

$$\langle \varphi'(\bar{x} + u), \bar{x} - \bar{y} \rangle = \langle \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle$$
$$+ \langle \varphi'(\bar{x} + u) - \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle$$
$$\geqslant \langle \varphi'(\bar{x}), \bar{x} - \bar{y} \rangle - \| \varphi'(\bar{x} + u) - \varphi'(\bar{x}) \| \| \bar{x} - \bar{y} \|$$
$$\geqslant \eta - L \| u \| \| \bar{x} - \bar{y} \| \geqslant L\delta \| \bar{x} - \bar{y} \|.$$

---

* It can easily be proved that if w is the center of a sphere R (of radius r) belonging to a regular set X, then we can put $r_1 = \frac{r\delta}{\| \bar{x} - w \| + r}$, the center of the sphere R_1 being at the point $w_1 = \frac{r_1}{r}(w - \bar{x})$, and an analogous construction applies to R_2.

The last three relations and the choice of δ prove Property C. $\triangle$

10.6.22. Definition. We say that an event $\mathfrak{M} = \mathfrak{M}(x, s)$ has occurred if there is $\tilde{\beta} > 0$ such that $y = x - \tilde{\beta}s \in U_\delta(\bar{y})$ where $x \in U_\delta(\bar{x})$ and $-s$ is a feasible direction found using the random descent method.

10.6.23. Property D. *If the event* $\mathfrak{M}_k = \mathfrak{M}(x_k, s_k)$ *occurs, then there is* $\varkappa > 0$ *such that*

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \varkappa.$$

Proof. Let $x_k \in U_\delta(\bar{x})$ and let y_k be an arbitrary point belonging to the intersection of the ray $x_k - \beta s_k$ and the set $U_\delta(\bar{y})$. Since $\omega_k = \inf_{0 \leqslant \beta \leqslant \zeta_0} \varphi(x_k - \beta s_k)$, for all $\beta \in [0, 1]$ there holds $\omega_k \leqslant \varphi(x_k - \beta(x_k - y_k))$. This and Lemma 9.2.3 imply the inequality

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k(\varphi(x_k) - \varphi(x_k - \beta(x_k - y_k)))$$

$$\geqslant \lambda_k\left(\beta\langle\varphi'(x_k), x_k - y_k\rangle - \frac{1}{2}L\beta^2\|x_k - y_k\|^2\right).$$

However, $\|x_k - y_k\| \leqslant 2\delta + \|\bar{x} - \bar{y}\| \equiv \theta$, and therefore Property C implies that for all $\beta \in [0, 1]$ we have

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \lambda_k\left(\frac{1}{2}\beta\eta - \frac{1}{2}L\beta^2\theta^2\right).$$

Let $\beta = \min\left\{1, \frac{1}{2}\frac{\eta}{L\theta^2}\right\}$. If $\beta = 1$, then $L\theta^2 \leqslant \frac{1}{2}\eta$ and

$$\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{1}{4}\lambda_k\eta.$$

If $\beta = \frac{1}{2}\frac{\eta}{L\theta^2}$, then $\varphi(x_k) - \varphi(x_{k+1}) \geqslant \frac{1}{8}\lambda_k\frac{\eta^2}{L\theta^2}$. Now, since $\lambda_k \geqslant \lambda > 0$, Property D takes place for $\varkappa = \frac{1}{4}\lambda\eta\min\left\{1, \frac{1}{2}\frac{\eta}{L\theta^2}\right\}$. $\triangle$

10.6.24. Property E. *If the sequence* $\{x_k\}$ *is such that* $x_k \in U_\delta(\bar{x})$, *then* $P\left\{\bigcup_{k=0}^{\infty}\mathfrak{M}_k\right\} = 1$.

Proof. Properties A and B imply that

$$\inf_{x \in U_\delta(\bar{x})} P\{\mathfrak{M} \mid x\} = p > 0,$$

For fixed $x_0, x_1, \ldots$ the corresponding events $\mathfrak{M}_0, \mathfrak{M}_1, \ldots$ are mutually independent since the random variables $s_0, s_1, \ldots$ are independent. Consequently,

$$P\left\{\bigcap_{k=0}^{m} \overline{\mathfrak{M}}_k \mid x_0, x_1, \ldots, x_m\right\} = \prod_{k=0}^{m} P\{\overline{\mathfrak{M}}_k \mid x_k\}$$

$$= \prod_{k=0}^{m} (1 - P\{\mathfrak{M}_k \mid x_k\}) \leqslant (1-p)^m \to 0, \quad m \to \infty.$$

Here $\overline{\mathfrak{M}}_k$ is the complement of the event $\mathfrak{M}_k$, i.e. the occurrence of $\overline{\mathfrak{M}}_k$ implies that the event $\mathfrak{M}_k$ does not occur. Therefore, according to the "zero-or-unity" rule, we have

$$P\left\{\bigcap_{k=0}^{\infty} \overline{\mathfrak{M}}_k \mid x_0, x_1, \ldots\right\} = 0,$$

$$\text{i.e. } P\left\{\bigcup_{k=0}^{\infty} \mathfrak{M}_k \mid x_0, x_1, \ldots\right\} = 1,$$

which implies that $P\left\{\bigcup_{k=0}^{\infty} \mathfrak{M}_k\right\} = 1$. $\triangle$

10.6.25. Theorem. $P\{\lim_{k\to\infty} \varphi(x_k) = \varphi^*\} = 1$.

Proof. By virtue of the assumption made in the procedure, the sequence of random variables $\{\varphi(x_k)\}$ is monotonic: $\varphi(x_{k+1}) \leqslant \varphi(x_k)$ $(k = 0, 1, \ldots)$ and bounded: $\varphi(x_k) \geqslant \varphi^*$ $(k = 0, 1, \ldots)$, and, consequently, there exists $\varphi = \lim_{k\to\infty} \varphi(x_k)$. Hence, with probability one there is an index k_0 such that for all $k = k_0 + 1, k_0 + 2, \ldots$ the inequality

$$\varphi(x_k) - \varphi(x_{k+1}) < \varkappa \tag{10.53}$$

holds.

If for any $\delta_1 > 0$, however small, there is a subsequence $\{x_{k_i}\}$ such that

$$x_{k_i} \in U_{\delta_1}(X^*) = \{x \in X: \rho(x, X^*) \leqslant \delta_1\},$$

then, with probability one, $\lim_{k\to\infty} \varphi(x_k) = \varphi^*$.

Assume that there is $\delta_1 > 0$ such that only a finite number of points belonging to the sequence $\{x_k\}$ fall in the neighborhood $U_{\delta_1}(X^*)$. We can assume, without loss

of generality, that $\mathbf{x}_k \notin U_{\delta_1}(X^*)$ for all k. Then, with probability one, there is a random index $k \geqslant k_0$ such that the event $\mathfrak{M}_k$ occurs, and, consequently, we have

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_{k+1}) > \varkappa.$$

This contradicts (10.53). The theorem is proved. △

Corollary. $\rho(\mathbf{x}_k, X^*) \xrightarrow[\text{a.c.}]{} 0$, $k \to \infty$.

This relation is an obvious consequence of the continuity of the function $\varphi(\mathbf{x})$, the boundedness of the set X_0, and the assertion of Theorem 10.6.25. △

10.6.26. The method of projecting a random direction. The minimization method presented in the foregoing sections may prove ineffective. The thing is that, as the dimension n of the space E_n increases, the probability that the chosen direction is feasible decreases. The rate of decrease can be characterized by a simple example where $X = \{\mathbf{x}: 0 \leqslant x_i \leqslant 1, \ i = \overline{1, n}\}$ and $\mathbf{x}_k = 0$. In this case the probability that a random direction $-\mathbf{s}_k$ is feasible is obviously equal to 2^{-n}. Therefore, generally, to find a feasible direction for large n, each iteration requires a greater number of trials, each trial involving the verification of whether the point $\mathbf{x}_k - \beta\mathbf{s}_k$ belongs to the set X for a sufficiently small β.

The present section is devoted to a method in which a random direction becomes feasible owing to the projection on the feasible region.

Denote by $\mathbf{z}_k$ the projection of the point $\mathbf{x}_k - \Delta_k\mathbf{r}_k$ on the set X. Here $\Delta_k = \operatorname{sgn}\langle\varphi'(\mathbf{x}_k), \mathbf{r}_k\rangle$ is a random vector uniformly distributed over an n-dimensional sphere of unit radius.

10.6.27. The scheme of the method.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k\mathbf{s}_k, \quad \mathbf{s}_k = \mathbf{x}_k - \mathbf{z}_k.$$

As β_k one takes an arbitrary number satisfying the conditions

$$\varphi(\mathbf{x}_k - \beta_k\mathbf{s}_k) \leqslant (1 - \lambda_k)\varphi(\mathbf{x}_k) + \lambda_k\omega_k,$$

$$0 \leqslant \beta_k \leqslant 1, \quad 0 < \lambda_k \leqslant \lambda_k \leqslant 1,$$

where $\omega_k = \min\limits_{\beta \in [0,1]} \varphi(\mathbf{x}_k - \beta\mathbf{s}_k)$.

10.6.28. The convergence of the method will be proved under the assumptions in Sec. 10.6.18; however the con-

dition of regularity is unnecessary here. To justify the convergence of the method, the same techniques as in Secs. 10.6.19-10.6.26 are applicable.

Consider the set* $V_\delta(\bar{y}) = \{y \in E_n: \| y - \bar{y} \| \leqslant \delta\}$. We note that $U_\delta(\bar{y}) = \{y \in X: \| y - \bar{y} \| \leqslant \delta\} \subseteq V_\delta(\bar{y})$.

We shall say that an event $\mathfrak{N} = \mathfrak{N}(x, r)$ has occurred if there is $\alpha > 0$ such that $x - \alpha r \in V_\delta(\bar{y})$ where $x \in U_\delta(\bar{x})$ and r is a direction found by the method of random search. The definition of the event $\mathfrak{M} = \mathfrak{M}(x, s)$ remains the same, however in this case $s = (x - y)/\| x - y \|$.

10.6.29. Property F. *If the event* $\mathfrak{N}_k = \mathfrak{N}(x_k, r_k)$ *occurs, then the event* $\mathfrak{M}_k = \mathfrak{M}(x_k, s_k)$ *also occurs.*

Proof. Let $v_k = x_k - \alpha r_k \in V_\delta(\bar{y})$ and let y_k be the projection of the point v_k on the set X. The properties of projections imply that $\| y_k - x \| \leqslant \| v_k - x \|$ for any $x \in X$, and, in particular, $\| y_k - \bar{y} \| \leqslant \| v_k - \bar{y} \|$, i.e. $y_k \in U_\delta(\bar{y})$. $\triangle$

Finally, it is obvious that for all $x \in U_\delta(\bar{x})$ there holds

$$P\{\mathfrak{M} \mid x\} \geqslant P\{\mathfrak{N} \mid x\} \geqslant \inf_{x \in U_\delta(\bar{x})} P\{\mathfrak{N} \mid x\} = p > 0.$$

One can now easily prove the convergence of the minimization process by repeating the argument used in the proof of Theorem 10.6.25.

---

* Here $\bar{y}$, $U_\delta(\bar{x})$, and $U_\delta(\bar{y})$ are the same as in Sec. 10.6.20.

Chapter 11

MODIFIED LAGRANGE FUNCTION METHOD

11.1. Lagrange Multiplier Method

11.1.1. The method consists in finding a saddle point $(\mathbf{x}^*, \mathbf{y}^*)$ of the Lagrange function $L(\mathbf{x}, \mathbf{y})$ on $E_n \times E_m^+$ (see Sec. 3.6). The construction of iterative processes is essentially based on the relation

$$L(\mathbf{x}^*, \mathbf{y}^*) = \max_{\mathbf{y} \in E_m^+} \min_{\mathbf{x} \in E_n} L(\mathbf{x}, \mathbf{y}) = \min_{\mathbf{x} \in E_n} \max_{\mathbf{y} \in E_m^+} L(\mathbf{x}, \mathbf{y}).$$

11.1.2. The first version of the method. This method suggested by Arrow and Hurwicz consists in constructing a sequence $\{(\mathbf{x}_k, \mathbf{y}_k)\}$ by means of the formulas

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta \frac{\partial}{\partial x} L(\mathbf{x}_k, \mathbf{y}_k),$$

$$\mathbf{y}_{k+1} = P_{E_m^+}\left(\mathbf{y}_k + \beta \frac{\partial}{\partial y} L(\mathbf{x}_k, \mathbf{y}_k)\right),$$

where $P_{E_m^+}(\mathbf{v})$ denotes the projection of the point $\mathbf{v}$ on E_m^+. Hence, the points $\mathbf{x}_k$ are constructed by the gradient descent method in application to the solution of the unconditional minimization problem for the function $L(\mathbf{x}, \mathbf{y})$ with respect to the variable $\mathbf{x}$, and the points $\mathbf{x}_k$ are found using a version of the gradient projection method to solve the problem of maximizing the function $L(\mathbf{x}, \mathbf{y})$ with respect to the variable $\mathbf{y}$ on the set $E_m^+ = \{\mathbf{y} \in E_m: \mathbf{y} \geqslant 0\}$.

The indicated process is advantageous due to its simplicity since

$$\frac{\partial}{\partial x} L(\mathbf{x}, \mathbf{y}) = \varphi'(\mathbf{x}) - \langle \mathbf{y}, f'(\mathbf{x}) \rangle,$$

$$\frac{\partial}{\partial y} L(\mathbf{x}, \mathbf{y}) = -f(\mathbf{x}), \quad f^T(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})),$$

and the projection on E_m^+ is performed quite easily.

Unfortunately, this method does not guarantee convergence even when both the gradient descent method and gradient projection method are sure to converge. The well-known example of the problem of one-dimensional minimization of x on the positive semiaxis $x \geqslant 0$ confirms what has been said.

11.1.3. The second version of the method. This version differs from the previous one in that the point x_{k+1} is found in each current iteration by means of total minimization of the function $L(x, y_k)$ with respect to the variable x:

$$x_{k+1} = \arg\min_{x \in E_n} L(x, y_k),$$
$$y_{k+1} = P_{E_m^+}(y_k - \beta f(x_k)).$$

In this version the convergence takes place in conditions under which Theorems 10.2.5 and 10.2.6 on convergence of the gradient projection procedure were proved.

11.2. Modified Lagrange Function Method

11.2.1. In the recent decade a great number of works have appeared on numerical methods of solving mathematical programming problems based on the construction of modified Lagrange functions. The modified Lagrange function method makes it possible to extend the area of application of the Lagrange multiplier method and is aimed at increasing the effectiveness of computational processes of finding saddle points.

The construction of a modified Lagrange function is based on the following ideas. For the problem of minimizing a function $\varphi(x)$ on a set X we introduce the function

$$N(x, y) = \varphi(x) - \bar{\gamma}(x, y)$$

defined throughout the space E_{n+m}.

The conditions on the function $\bar{\gamma}(x, y)$ are stated in the following theorem.

11.2.2. Theorem. *If there are points* $x^* \in X$ *and* $y^* \in E_m$ *such that*

(1) $x^* \in X(y^*) = \operatorname{Arg\,min}\{N(x, y^*): x \in E_n\}$,

(2) $\bar{\gamma}(x, y^*) \geqslant \bar{\gamma}(x^*, y^*)$ $\forall x \in X$,

then

$$\mathbf{x}^* \in X^* = \operatorname{Arg\,min} \{\varphi(\mathbf{x}): \mathbf{x} \in X\} = X \cap X(\mathbf{y}^*).$$

Proof. Since $\mathbf{x}^* \in X \cap X(\mathbf{y}^*)$, we have

$$N(\mathbf{x}^*, \mathbf{y}^*) = \varphi(\mathbf{x}^*) - \bar{\gamma}(\mathbf{x}^*, \mathbf{y}^*) \leqslant \varphi(\mathbf{x}) - \bar{\gamma}(\mathbf{x}, \mathbf{y}^*)$$

for all $\mathbf{x} \in E_n$. In particular, in view of the condition (2), for $\mathbf{x} \in X$ we obtain $\varphi(\mathbf{x}^*) \leqslant \varphi(\mathbf{x})$ $\forall \mathbf{x} \in X$, and, consequently, $\mathbf{x}^* \in X^*$. Therefore $X \cap X(\mathbf{y}^*) \subset X^*$.

We now show that $X^* \setminus (X \cap X(\mathbf{y}^*)) = \varnothing$, which completes the proof of the theorem.

Assume that there is $\bar{\mathbf{x}} \in X^*$ such that $\bar{\mathbf{x}} \notin X \cap X(\mathbf{y}^*)$. Since $\bar{\mathbf{x}} \in X^* \subset X$, the assumption implies that $\bar{\mathbf{x}} \notin X(\mathbf{y}^*)$, and, consequently, $N(\mathbf{x}^*, \mathbf{y}^*) = \varphi(\mathbf{x}^*) - \bar{\gamma}(\mathbf{x}^*, \mathbf{y}^*) < \varphi(\bar{\mathbf{x}}) - \bar{\gamma}(\bar{\mathbf{x}}, \mathbf{y}^*) \leqslant \varphi(\bar{\mathbf{x}}) - \bar{\gamma}(\mathbf{x}^*, \mathbf{y}^*)$. We thus obtain the inequality $\varphi(\mathbf{x}^*) < \varphi(\bar{\mathbf{x}})$ contradicting the assumption. $\triangle$

11.2.3. The modified Lagrange function. We consider the basic mathematical programming problem

$$\varphi(\mathbf{x}) \to \min, \quad \mathbf{x} \in X = \{\mathbf{x}: f_i(\mathbf{x}) \geqslant 0, \ i = \overline{1, m}\}. \tag{11.1}$$

As is known, the function

$$L(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \sum_{i=1}^{m} y_i f_i(\mathbf{x}) \tag{11.2}$$

defined for all $\mathbf{x} \in E_n$ and $\mathbf{y} \in E_m^+ = \{\mathbf{y}: \mathbf{y} \geqslant 0\}$ is the Lagrange function for the problem (11.1).

Consider the function

$$N(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{x}) - \sum_{i=1}^{m} \gamma(f_i(\mathbf{x}), y_i) \tag{11.3}$$

and denote

$$\bar{\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \gamma(f_i(\mathbf{x}), y_i).$$

If the function $\gamma(f_i(\mathbf{x}), y_i)$ $(i = \overline{1, m})$ is such that the conditions of Theorem 11.2.2 are fulfilled, then $N(\mathbf{x}, \mathbf{y})$ will be called the modified Lagrange function. Thus, if the function $\bar{\gamma}(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{m} \gamma(f_i(\mathbf{x}), y_i)$ is

such that there exist $\mathbf{x}^*$ and $\mathbf{y}^*$ satisfying the conditions

$$\begin{aligned} &\mathbf{x}^* \in X \cap X(\mathbf{y}^*), \\ &\bar{\gamma}(\mathbf{x}, \mathbf{y}^*) \geqslant \bar{\gamma}(\mathbf{x}^*, \mathbf{y}^*) \quad \forall \mathbf{x} \in X, \end{aligned} \tag{11.4}$$

then $\mathbf{x}^* \in X$, and $N(\mathbf{x}, \mathbf{y})$ is the modified Lagrange function for the problem (11.1)

11.2.4. Constructing the modified Lagrange function. As is known, the equation

$$\frac{\partial}{\partial x} L(\mathbf{x}, \mathbf{y}) = \nabla\varphi(\mathbf{x}) - \sum_{i=1}^{m} y_i \nabla f_i(\mathbf{x}) = \mathbf{0} \tag{11.5}$$

determines a point $\mathbf{x}$ where the necessary conditions for the function $L(\mathbf{x}, \mathbf{y})$ to have minimum with respect to $\mathbf{x}$ are fulfilled. For the function $N(\mathbf{x}, \mathbf{y})$ the necessary condition for minimum with respect to $\mathbf{x}$ is written

$$\frac{\partial}{\partial x} N(\mathbf{x}, \mathbf{y}) = \nabla\varphi(\mathbf{x}) - \sum_{i=1}^{m} \frac{\partial}{\partial f_i} \gamma(f_i(\mathbf{x}), y_i) \nabla f_i(\mathbf{x}) = \mathbf{0}. \tag{11.6}$$

Comparison of (11.5) and (11.6) suggests the idea to introduce the system of equations

$$y_i = \frac{\partial}{\partial f_i} \gamma(f_i(\mathbf{x}), y_i), \quad i = \overline{1, m}, \tag{11.7}$$

and to require its solvability under the additional conditions

$$f_i(\mathbf{x})\, y_i = 0, \quad i = \overline{1, m}, \quad \forall \mathbf{x} \in X. \tag{11.8}$$

11.2.5. Conditions for the existence of the function $N(x, y)$. We shall consider conditions under which the modified Lagrange function can be constructed under rather general assumptions on the properties of the functions $f_i(\mathbf{x})$ $(i = \overline{1, m})$.

(A) If $f_i(\mathbf{x}) < 0$ for an index i, then the ith equation in (11.7) has no solution, and, consequently the solution set of the system (11.7) is empty;

if $f_i(\mathbf{x}) > 0$ for an index i, then the only solution of the ith equation in the system (11.7) is $y_i = 0$;

if $f_i(\mathbf{x}) = 0$ for an index i, then any $y_i \geqslant 0$ is a solution of the ith equation in the system (11.7).

(B) If for a pair $\mathbf{x}^*$, $\mathbf{y}^* \geqslant 0$ the conditions (11.8) hold, then

$$\sum_{i=1}^{m} \gamma(f_i(\mathbf{x}), y_i^*) \geqslant \sum_{i=1}^{m} \gamma(f_i(\mathbf{x}^*), y_i^*) \quad \forall \mathbf{x} \in X.$$

11.2.6. Theorem. *If the functions* $\gamma(f_i(\mathbf{x}), y_i)$ $(i = \overline{1, m})$ *are such that*

(a) *the conditions* (A) *are fulfilled;*

(b) *there are* $\mathbf{x}^* = \mathbf{x}(\mathbf{y}^*) \in X(\mathbf{y}^*)$ *and* $\mathbf{y}^* \geqslant 0$ *such that*

$$\frac{\partial}{\partial f_i} \gamma(f_i(\mathbf{x}^*), y_i^*) = y_i^*, \quad i = \overline{1, m};$$

(c) *for* $\mathbf{x}^* = \mathbf{x}(\mathbf{y}^*)$ *and* $\mathbf{y}^* \geqslant 0$ *the conditions* (B) *hold, then* $\mathbf{x}^* \in X^*$.

Moreover, if the function $N(\mathbf{x}, \mathbf{y}^*)$ *is differentiable at the point* $\mathbf{x}^*$ *and for any* $f_i(\mathbf{x})$ *and* $y_i \geqslant 0$ *there holds*

$$\frac{\partial}{\partial f_i} \gamma(f_i(\mathbf{x}), y_i) \leqslant 0, \quad i = \overline{1, m}, \tag{11.9}$$

then the pair $\mathbf{x}^*$, $\mathbf{y}^*$ *is a Kuhn-Tucker point.*

Proof. The conditions of the theorem imply that there are $\mathbf{x}^*$, $\mathbf{y}^*$ such that

$$f_i(\mathbf{x}^*) \geqslant 0, \quad y_i^* \geqslant 0 \quad \text{and} \quad f_i(\mathbf{x}^*)\, y_i^* = 0, \quad i = \overline{1, m},$$

and the inequality (11.4) holds. Consequently, all the conditions of Theorem 11.2.2 are fulfilled, and hence $\mathbf{x}^* \in X^*$.

Taking into account (11.7) and (11.9) we conclude that the pair $\mathbf{x}^*$, $\mathbf{y}^*$ is a Kuhn-Tucker point.

11.2.7. Two examples of a function $\gamma(f_i(\mathbf{x}), y_i)$:

$$\gamma_1(f_i(\mathbf{x}), y_i) = -\frac{1}{2}[\min\{f_i(\mathbf{x}) - y_i, 0\}]^2,$$

$$\gamma_2(f_i(\mathbf{x}), y_i) = \frac{1}{2} y_i^2 - \frac{1}{2}[\min\{f_i(\mathbf{x}) - y_i, 0\}]^2.$$

It can easily be verified that the function $-\gamma_1$ is convex with respect to $\mathbf{x}$ if the function $f_i(\mathbf{x})$ is concave

and is convex with respect to $\mathbf{y}$ (see Sec. 2.4.8). The function $-\gamma_2$ is convex with respect to $\mathbf{x}$ and concave with respect to $\mathbf{y}$. The latter property follows from the fact that for $f_i(\mathbf{x}) - y_i < 0$ $-\gamma_2$ is linear in $\mathbf{y}$: $-\gamma_2(f_i(\mathbf{x}), y_i) = -\frac{1}{2}y_i^2 + \frac{1}{2}(f_i(\mathbf{x}) - y_i)^2 = -f_i(\mathbf{x})\, y_i + \frac{1}{2}f_i^2(\mathbf{x})$ and for $f_i(\mathbf{x}) - y_i \geqslant 0$ it is strictly concave: $-\gamma_2(f_i(\mathbf{x}), y_i) = -\frac{1}{2}y_i^2$.

It suffices to verify the properties (A) and (B) for $\gamma_2(f_i(\mathbf{x}), y_i)$:

$$\frac{\partial}{\partial f_i}\gamma_2(f_i(\mathbf{x}), y_i) = \begin{cases} y_i - f_i(x) & \text{if } y_i \geqslant f_i(\mathbf{x}), \\ 0 & \text{if } y_i < f_i(\mathbf{x}). \end{cases}$$

(A) If $f_i(\mathbf{x}) < 0$ and $y_i \geqslant f_i(\mathbf{x})$, then (11.7) implies $y_i - f_i(\mathbf{x}) = y_i$, whence $f_i(\mathbf{x}) = 0$. This contradiction indicates that the system (11.7) is unsolvable. If $f_i(\mathbf{x}) > y_i$, then $y_i < 0$, and, since $\frac{\partial}{\partial f_i}\gamma_2(f_i(\mathbf{x}), y_i) = 0$, we see that (11.7) implies $y_i = 0$, which is a contradiction.

If $f_i(\mathbf{x}) > 0$ and $y_i \geqslant f_i(\mathbf{x})$, we again arrive at a contradiction. If $f_i(\mathbf{x}) > y_i$, then $\frac{\partial}{\partial f_i}\gamma_2(f_i(\mathbf{x}), y_i) = 0$, and, consequently, the only solution to the ith equation is $y_i = 0$. If $f_i(\mathbf{x}) = 0$ and $y_i \geqslant f_i(\mathbf{x})$, then $y_i \geqslant 0$ and $\frac{\partial}{\partial f_i}\gamma_2(f_i(\mathbf{x}), y_i) = y_i$, and therefore any $y_i \geqslant 0$ satisfies the ith equation in the system (11.7). The case $y_i < f_i(\mathbf{x}) = 0$ is contradictory since (11.7) implies that $y_i = 0$.

(B) If $f_i(\mathbf{x}) - y_i^* \geqslant 0$, then $\gamma_2(f_i(\mathbf{x}), y_i^*) = \frac{1}{2}(y_i^*)^2 \geqslant \gamma_2(f_i(\mathbf{x}^*), y_i^*)$. If $f_i(\mathbf{x}) - y_i^* < 0$, $\mathbf{x} \in X$, then $y_i^* > f_i(\mathbf{x}) \geqslant 0$ and $f_i(\mathbf{x}^*) = 0$. Since $-2y_i^* f_i(\mathbf{x}) + (f_i(\mathbf{x}))^2 < 0$ and $y_i^* f_i(\mathbf{x}^*) = 0$, we have

$$[\min\{f_i(\mathbf{x}) - y_i^*, 0\}]^2 = (f_i(\mathbf{x}) - y_i^*)^2 - (y_i^*)^2 - 2y_i^* f_i(\mathbf{x}) + (f_i(\mathbf{x}))^2 < (y_i^*)^2 = (y_i^* - f_i(\mathbf{x}^*))^2 = [\min\{f_i(\mathbf{x}^*) - y_i^*; 0\}]^2,$$

and, consequently $\gamma_2(f_i(\mathbf{x}), y_i^*) \geqslant \gamma_2(f_i(\mathbf{x}^*), y_i^*)$.

11.2.8. Discussion. In conditions of Theorem 11.2.2 the pair $(\mathbf{x}^*, \mathbf{y}^*)$ may not be a Kuhn-Tucker point. For instance, solving the problem (11.1) by the penalty

function method with penalty

$$\frac{1}{\beta}\psi(x) = \frac{1}{\beta}\sum_{i=1}^{m}[\min\{f_i(x), 0\}]$$

for sufficiently small $\beta \in (0, \beta_0]$ we obtain

$$\min_{x\in X}\varphi(x) = \min_{x\in E_n}\left(\varphi(x) + \frac{1}{\beta}\psi(x)\right)$$

if the function $\varphi(x)$ is strongly convex and the functions $f_i(x)$ are 1-majorizable ($\gamma = 1$) and continuously differentiable (see Sec. 6.2.7). If we put $y_i = 1/\beta$ ($i = \overline{1, m}$) and $\bar{\gamma}(x, y) = -\frac{1}{\beta}\psi(x)$, then at the solution point x^* the conditions (1) and (2) of Theorem 11.2.2 are fulfilled for $y^* = (1/\beta^*, \ldots, 1/\beta^*)$, where $\beta^* \in (0, \beta_0]$. However, in the general case of setting the functions $f_i(x)$ the condition that all the components of the vector y^* are equal is sure to be violated for Kuhn-Tucker points.

The fact that a pair (x^*, y^*) is a Kuhn-Tucker point is guaranteed by the conditions of Theorem 11.2.6. At the same time, the function $N(x, y)$ itself may not have saddle points despite the fact that the original mathematical programming problem is convex. For example, the function $N(x, y) = \varphi(x) - \bar{\gamma}_1(x, y)$ satisfies all the conditions of Theorem 11.2.6 but has no saddle point. A natural condition guaranteeing the existence of a saddle point of the function $N(x, y)$ is the inequality $\bar{\gamma}(x^*, y) \geqslant \bar{\gamma}(x^*, y^*)$ for all $y \in E_m$. For instance, the function $\sum_{i=1}^{m}\gamma_2(f_i(x), y_i)$ possesses this property.

Numerical realization. One way is to find the point x^* by solving the system of equations (11.6), (11.7), for instance, with the aid of the following iterative process:

$$x_k \in X(y_k) = \operatorname{Arg\,min}\{N(x, y_k): x \in E_n\},$$

$$y_{k+1} = \frac{\partial}{\partial f}\bar{\gamma}(x_k, y_k).$$

The application of gradient methods for solving the system (11.6), (11.7) involves the calculation of second deriva-

tives of the function $\bar{\gamma}(\mathbf{x}, \mathbf{y})$. However, even when the functions $\varphi(\mathbf{x})$ and $f_i(\mathbf{x})$ are twice differentiable, the second derivatives of many of the well-known modified Lagrange functions are not defined everywhere as is required when using gradient methods to solve the equations (11.7). This very situation takes place in the above examples. In such cases one should resort to methods of minimizing non-differentiable functions.

The other way is to calculate the point $\mathbf{x}^*$ by means of methods of finding saddle points in application to the function $N(\mathbf{x}, \mathbf{y})$ provided that such points exist.

In this case first-order methods are applicable, for instance, the following analog of the Arrow-Hurwicz method:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta \frac{\partial}{\partial x} N(\mathbf{x}_k, \mathbf{y}_k),$$

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \beta \frac{\partial}{\partial y} N(\mathbf{x}_k, \mathbf{y}_k).$$

11.3. Interrelation between Lagrange Multiplier Method and Penalty Function Method

11.3.1. As was seen, the basic idea of the method of modified Lagrange functions and the penalty function method consists in reducing optimization problems with constraints to unconditional problems. Theorem 11.2.6 established conditions under which the solution to the problem obtained by the modified Lagrange function method is a Kuhn-Tucker point. It turns out that the penalty function method also possesses this property if the partial derivatives of the penalty with respect to the constraint functions are taken as dual variables.

We shall find out conditions under which the penalty function method also leads to a Kuhn-Tucker point. In this way the interrelation between the modified Lagrange function and penalty function methods will be established.

11.3.2. Consider the basic mathematical programming problem

$$\varphi(\mathbf{x}) \to \min, \quad \mathbf{x} \in X = \{\mathbf{x}: f(\mathbf{x}) \geqslant 0\}, \quad (11.10)$$

where $f^T(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x}))$.

We write the penalty function in the following form:

$$M(x, \beta) = \varphi(\mathbf{x}) + \frac{1}{\beta}\psi(f(\mathbf{x})), \quad \beta > 0,$$

where

$\psi(f(\mathbf{x})) = 0$ if $f_i(\mathbf{x}) \geqslant 0$ for all $i \in I = \{1, 2, \ldots, m\}$,
$\psi(f(\mathbf{x})) > 0$ if $f_i(\mathbf{x}) < 0$ for at least one $i \in I$.

Notation. $L(\mathbf{x}, \mathbf{u}) = \varphi(\mathbf{x}) + \langle \mathbf{u}, f(\mathbf{x}) \rangle$ ($\mathbf{u}^T = (u_1, u_2, \ldots, u_m) \leqslant 0$) is the Lagrange function for the problem (11.1); $Y = \operatorname{Arg\,min}\{\varphi(\mathbf{x}): \mathbf{x} \in X\}$;

$$Y_\beta = \Big\{\mathbf{x}: \frac{\partial}{\partial x} M(\mathbf{x}, \beta) = \nabla\varphi(\mathbf{x}) + \frac{1}{\beta}\sum_{i=1}^{m}\Big(\frac{\partial}{\partial f_i}\psi(f(\mathbf{x}))\Big)\nabla f_i(\mathbf{x}) = 0,\ \beta > 0\Big\}$$

is the set of stationary points of the penalty function;

$$K = \Big\{(\mathbf{x}, \mathbf{u}): \frac{\partial}{\partial x} L(\mathbf{x}, \mathbf{u}) = \nabla\varphi(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x}) = 0;\ u_i f_i(\mathbf{x}) = 0,\ \mathbf{u} \leqslant 0\Big\}$$

is the set of Kuhn-Tucker points;

$$u_i(\beta) = \frac{1}{\beta}\frac{\partial}{\partial f_i}\psi(f(\mathbf{y}_\beta)); \quad i = \overline{1, m}, \quad \mathbf{y}_\beta \in Y_\beta,$$

$$\mathbf{v}(\beta) = (\mathbf{y}_\beta, \mathbf{u}(\beta));$$

$$\mathbf{u}^T(\beta) = (u_1(\beta), u_2(\beta), \ldots, u_m(\beta)), \quad I(\mathbf{x}) = \{i: f_i(\mathbf{x}) = 0\}.$$

11.3.3. Assumptions. (1) $\varphi(\mathbf{x})$ is a convex function, $f_i(\mathbf{x})$, $i = \overline{1, m}$, are concave functions, and $\psi(f(\mathbf{x}))$ is a convex function.

(2) The set X is regular in Slater's sense.

(3) The set Y is nonempty and bounded.

(4) $\psi(f(\mathbf{x})) \in C^1(E_n)$.

(5) $\frac{\partial}{\partial f_i}\psi(f(\mathbf{x})) \leqslant 0$, $i = \overline{1, m}$.

(6) $\frac{\partial}{\partial f_i}\psi(f(\mathbf{x})) = 0$ if $0 < f_i(\mathbf{x}) \leqslant \mu < \infty$.

11.3.4. Theorem. *If the conditions* (1) *to* (6) *hold, then*

$$\lim_{\beta \to 0} \rho(\mathbf{v}(\beta), K) = 0.$$

Proof. Let $\mathbf{y}_\beta \in Y_\beta$, then

$$\nabla\varphi(\mathbf{y}_\beta) = \sum_{i=1}^{m} -u_i(\beta)\nabla f_i(\mathbf{y}_\beta).$$

If we prove that the family $\{\mathbf{v}(\beta)\}$ is uniformly bounded with respect to β, this would imply that any limiting point $\mathbf{v}^* = (\mathbf{y}^*, \mathbf{u}^*)$ satisfies the relation $\nabla\varphi(\mathbf{y}^*) + \sum_{i=1}^{m} u_i^* \nabla f_i(\mathbf{y}^*) = 0$, and, moreover, by virtue of the condition (6), $u_i^* f_i(\mathbf{y}^*) = 0$. Thus, any limiting point $\mathbf{v}^*$ will belong to the set K, and, consequently, $\rho(\mathbf{v}(\beta), K) \to 0$ $(\beta \to 0)$.

Assume that the family $\{\mathbf{u}(\beta)\}$ is not uniformly bounded with respect to β. Then there is a sequence $\{\beta_k\}$ such that $-u_i(\beta_k) \to +\infty$, $\beta_k \to 0$ for at least one index i. The assumption that the set Y is bounded implies that the sets Y_β are bounded jointly for all β (see Theorem 1 in Suppl. 4). This theorem implies that any limiting point $\mathbf{y}^*$ of the family $\{\mathbf{y}_\beta\}$ belongs to the set Y.

We can assume, without loss of generality, that $\lim_{k \to \infty} \mathbf{y}_{\beta_k} = \mathbf{y}^*$. By the condition (2), there is a point $\mathbf{z} \in X$ such that $f_i(\mathbf{z}) > 0$ $(i = \overline{1, m})$. The convexity of the function $\varphi(\mathbf{x})$ implies

$$\varphi(\mathbf{z}) - \varphi(\mathbf{y}_\beta) \geqslant \langle \nabla\varphi(\mathbf{y}_\beta), \mathbf{z} - \mathbf{y}_\beta \rangle = \sum_{i=1}^{m} -u_i(\beta) \langle \nabla f_i(\mathbf{y}_\beta), \mathbf{z} - \mathbf{y}_\beta \rangle.$$

Since $f_i(\mathbf{x})$ are concave functions, we have $f_i(\mathbf{z}) - f_i(\mathbf{y}_\beta) \leqslant \langle \nabla f_i(\mathbf{y}_\beta), \mathbf{z} - \mathbf{y}_\beta \rangle$, and taking into consideration that $-u_i(\beta) \geqslant 0$ $(i = \overline{1, m})$, we find

$$\varphi(\mathbf{z}) - \varphi(\mathbf{y}_\beta) \geqslant \sum_{i=1}^{m} -u_i(\beta)(f_i(\mathbf{z}) - f_i(\mathbf{y}_\beta)).$$

Assume that $I(\mathbf{y}^*) \neq \varnothing$. Then there is an index $j \in I(\mathbf{y}^*)$ such that $f_j(\mathbf{y}^*) = 0$, and for sufficiently small $\beta_k > 0$

there must be $f_j(\mathbf{y}_{\beta_k}) < \frac{1}{2} f_j(\mathbf{z})$. Hence, $f_j(\mathbf{z}) - f_j(\mathbf{y}_{\beta_k}) \leqslant \frac{1}{2} f_j(\mathbf{z})$ for $j \in I(\mathbf{y}^*)$ and sufficiently large indices k. From the conditions (5) and (6) we conclude that there is an index k_0 such that $u_i(\beta_k) = 0 \ \forall i \in I \setminus I(\mathbf{y}^*)$ for all $k \geqslant k_0$. This follows from the fact that $f_i(\mathbf{y}^*) > 0$ for $i \in I \setminus I(\mathbf{y}^*)$, and, since $\mathbf{y}_{\beta_k} \to \mathbf{y}^*$, we have $f_i(\mathbf{y}_{\beta_k}) > 0$ for $k \geqslant k_0$.

Further,

$$\varphi(\mathbf{z}) - \varphi(\mathbf{y}_{\beta_k}) \geqslant \sum_{i=1}^{m} -u_i(\beta_k)\,(f_i(\mathbf{z}) - f_i(\mathbf{y}_{\beta_k}))$$

$$= \sum_{i \in I(\mathbf{y}^*)} -u_i(\beta_k)\,(f_i(\mathbf{z}) - f_i(\mathbf{y}_{\beta_k}))$$

$$> \frac{1}{2} \sum_{i \in I(\mathbf{y}^*)} -u_i(\beta_k)\, f_i(\mathbf{z}),$$

and therefore

$$\sum_{i \in I(\mathbf{y}^*)} -u_i(\beta_k) \leqslant 2\,\frac{\varphi(\mathbf{z}) - \varphi(\mathbf{y}_{\beta_k})}{\min\limits_{i \in I(\mathbf{y}^*)} f_i(\mathbf{z})} \leqslant 2\,\frac{\varphi(\mathbf{z}) - \varphi(\mathbf{y}_{\beta_{k_0}})}{\min\limits_{i \in I(\mathbf{y}^*)} f_i(\mathbf{z})} = \text{const}.$$

The right-hand inequality is implied by the fact that the sequence $\{\varphi(\mathbf{y}_{\beta_k})\}$ is monotonically increasing (see (6.8) in Sec. 6.2.2).

We have thus arrived at a relation contradicting the assumption that the sequence $\{-u_i(\beta_k)\}$ is unbounded.

Finally, in view of the condition (6), the case when $I(\mathbf{y}^*) = \varnothing$ is trivial.

SUPPLEMENTS

Supplement 1. On Finiteness of the Numerical Realization of the Method for Determining Steplength from Condition (9.15)

We can assert that for functions of class $C^{1,1}$ (E_n) the determination of the coefficient β_k from the condition (9.15) is realized after a finite number of duplications (or subdivisions) of the steplength.

Thus, we put $\beta_k = \dfrac{\partial \varphi(\mathbf{x}_k)}{\partial \mathbf{s}_k} 2^i$ $(i = 0, \pm 1, \pm 2, \ldots)$, $\|\mathbf{s}_k\| = 1$, and verify the fulfilment of (9.15). The index i varies until the inequality (9.15) is fulfilled. Let (9.15) be first fulfilled for $i = i_0(k)$. Then we put $\beta_k = \dfrac{\partial \varphi(\mathbf{x}_k)}{\partial \mathbf{s}_k} 2^{i_0(k)}$.

We shall show that the indices $i_0(k)$ are bounded jointly, i.e. $|i_0(k)| \leqslant M$ $(k = 0, 1, \ldots)$. Indeed, for $i \leqslant N = [-\log_2 L] - 1$ we have $(1 - L2^{i-1}) \geqslant 1/2$ (here $[a]$ denotes the integral part of the number a), whence

$$\varphi(\mathbf{x}_k) - \varphi\left(\mathbf{x}_k - 2^i \frac{\partial \varphi(\mathbf{x}_k)}{\partial \mathbf{s}_k} \mathbf{s}_k\right) \geqslant \langle \varphi'(\mathbf{x}_k), 2^i \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle \mathbf{s}_k \rangle$$

$$- \frac{L 2^{2i}}{2} \|\langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle\|^2 = (2^i - 2L^{2i-1}) \left\| \frac{\partial \varphi(\mathbf{x}_k)}{\partial \mathbf{s}_k} \right\|^2$$

$$\geqslant \frac{1}{2} \frac{\partial \varphi(\mathbf{x}_k) 2^i}{\partial \mathbf{s}_k} \left\| \frac{\partial \varphi(\mathbf{x}_k)}{\partial \mathbf{s}_k} \right\|, \quad k \in I = \{0, 1, 2, \ldots\}.$$

Further, this means that the inequality (9.15) holds for $\beta_k = 2^{N-1} \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle$ as well. Consequently, the index $i_0(k)$ for which (9.15) is fulfilled does not exceed $M = |N - 1|$, and the coefficient $\beta_k = 2^{i_0(k)} \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle$ satisfies (9.15). We note that β_k can be represented as $\beta_k = \dfrac{C_k}{L} \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle$ where $L 2^{-|N-1|} \leqslant C_k \leqslant L 2^{|N-1|}$. Therefore the inequality (9.15) is fulfilled after no more than a finite number of duplications carried out uniformly over all the iterations.

In the case of a strongly convex function $\varphi(\mathbf{x})$ the determination of the coefficient β_k from the condition

$$\beta_k = \arg \min_{\beta \geqslant 0} \varphi(\mathbf{x} - \beta \mathbf{s}_k)$$

can be replaced by the following estimate of this coefficient from below and from above:

$$\langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle/L \leqslant \beta_k \leqslant \langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle/\rho, \tag{1}$$

where ρ is the parameter of strong convexity. Indeed,

$$\rho\beta_k^2 \leqslant \langle \varphi'(\mathbf{x}_k),\ \beta_k\mathbf{s}_k\rangle = \beta_k \langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle,$$

and, consequently, $\beta_k \leqslant \langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle/\rho$. On the other hand,

$$\langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle = \langle \varphi'(\mathbf{x}_k) - \varphi'(\mathbf{x}_k - \beta_k\mathbf{s}_k),\ \mathbf{s}_k\rangle$$
$$\leqslant \|\varphi'(\mathbf{x}_k) - \varphi'(\mathbf{x}_k - \beta_k\mathbf{s}_k)\| \leqslant L\beta_k,$$

and therefore $\beta_k \geqslant \langle \varphi'(\mathbf{x}_k),\ \mathbf{s}_k\rangle/L$, which, together with the foregoing inequality, implies (1).

Supplement 2. Estimating the Rate of Convergence for Cyclic Method of Coordinate-Wise Descent with Duplication of Steplength

In this method the direction of descent is the same as in the case 9.7.8, and the steplength β_k is found by way of duplication according to the scheme described in 9.4.6, i.e. from the condition that the inequality

$$\varphi(\mathbf{x}_k) - \varphi(\mathbf{x}_k - \beta_k\mathbf{s}_k) \geqslant \frac{\beta_k}{2}\,\frac{\partial\varphi(\mathbf{x}_k)}{\partial\mathbf{s}_k}$$

is fulfilled.

We shall show that for this method a linear rate of convergence with respect to the functional takes place.

Lemma. *For any vector* $\mathbf{h} \in E_n$, $\|\mathbf{h}\| = 1$, *there is an index* $i \in \{k, \ldots, k-n-1\}$ *such that*

$$|\langle \mathbf{h},\ \mathbf{s}_i\rangle| \geqslant 1/\sqrt{n}, \qquad k = 0,\ 1,\ \ldots \tag{1}$$

Proof. It follows from (1) that there is at least one component of the vector $\mathbf{h}$ such that $|h_j| \geqslant 1/\sqrt{n}$, and, consequently, for the vector $\mathbf{s}_i = \mathbf{e}_j$ the inequality (1) is fulfilled. $\triangle$

Corollary. *For any* $\mathbf{h} \in E_n$, $\|\mathbf{h}\| = 1$, *there is a neighborhood* $U_\delta(\mathbf{h}) = \{\mathbf{x}\colon \rho(\mathbf{x}, \mathbf{h}) \leqslant \delta\}$ *such that* $\langle \mathbf{x}, \mathbf{s}_i\rangle \leqslant 1/(2\sqrt{n})$ *for all* $\mathbf{x} \in U_\delta(\mathbf{h})$, $\delta \leqslant 1/(2\sqrt{n})$, *and some* $i \in \{k, \ldots, k+n-1\}$, $k = 0,\ 1,\ \ldots$.

Proof. Let $\mathbf{x} \in U_\delta(\mathbf{h})$. Then $\mathbf{x} = \mathbf{h} + \mathbf{y}$ where $\|\mathbf{y}\| \leqslant 1/(2\sqrt{n})$ and, consequently, $\langle \mathbf{x}, \mathbf{s}_i\rangle = \langle \mathbf{h}, \mathbf{s}_i\rangle + \langle \mathbf{y}, \mathbf{s}_i\rangle \geqslant 1/(2\sqrt{n})$. $\triangle$

Theorem 1. *Let the sequence* $\{\mathbf{x}_k\}$ *be constructed according to the scheme* 9.7.8 *and let* $\beta_k = \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k\rangle/L$. *Then for any index* p *and at least one index* $i \in \{0, 1, \ldots, n-1\}$ *the inequality*

$$\alpha_{p+i} = \langle \varphi'(\mathbf{x}_{p+i}),\ \mathbf{s}_{p+i}\rangle/\|\varphi'(\mathbf{x}_{p+i})\| \geqslant 1/(4n\sqrt{n}) \tag{2}$$

holds.

Proof. It follows from the corollary that for any $p = 0, 1, \ldots$ there exists s_{p+i_0} such that

$$\langle x, s_{p+i_0}\rangle \geqslant 1/(2\sqrt{n}), \quad x \in U_\delta(\varphi'(x_p)/\|\varphi'(x_p)\|),$$

where $\delta = 1/(2\sqrt{n})$, $i_0 < n$. For any $i \in \{0, \ldots, n-1\}$ we have

$$\left\| \frac{\varphi'(x_p)}{\|\varphi'(x_p)\|} - \frac{\varphi'(x_{p+i})}{\|\varphi'(x_{p+i})\|} \right\| = \left\| \frac{\varphi'(x_p)}{\|\varphi'(x_p)\|} - \frac{\varphi'(x_{p+1})}{\|\varphi'(x_{p+1})\|} + \frac{\varphi'(x_{p+1})}{\|\varphi'(x_{p+1})\|} - \frac{\varphi'(x_{p+2})}{\|\varphi'(x_{p+2})\|} + \ldots + \frac{\varphi'(x_{p+i-1})}{\|\varphi'(x_{p+i-1})\|} - \frac{\varphi'(x_{p+i})}{\|\varphi'(x_{p+i})\|} \right\| \leqslant \left\| \frac{\varphi'(x_p)}{\|\varphi'(x_p)\|} - \frac{\varphi'(x_{p+1})}{\|\varphi'(x_{p+1})\|} \right\| + \ldots + \left\| \frac{\varphi'(x_{p+i-1})}{\|\varphi'(x_{p+i-1})\|} - \frac{\varphi'(x_{p+i})}{\|\varphi'(x_{p+i})\|} \right\|. \quad (3)$$

We estimate the jth term on the right-hand side of (3):

$$\left\| \frac{\varphi'(x_{p+j})}{\|\varphi'(x_{p+j})\|} - \frac{\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j+1})\|} \right\| = \left\| \frac{\|\varphi'(x_{p+j+1})\|\,\varphi'(x_{p+j}) - \|\varphi'(x_{p+j})\|\,\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j+1})\|\,\|\varphi'(x_{p+j})\|} \right\|.$$

Adding $\pm\|\varphi'(x_{p+j+1})\|\,\varphi'(x_{p+j+1})$ to the numerator in the right-hand side of the above relation we find

$$\left\| \frac{\|\varphi'(x_{p+j+1})\|\,\varphi'(x_{p+j}) - \|\varphi'(x_{p+j+1})\|\,\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j})\|\,\|\varphi'(x_{p+j+1})\|} + \frac{\|\varphi'(x_{p+j+1})\|\,\varphi'(x_{p+j+1}) - \|\varphi'(x_{p+j})\|\,\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j})\|\,\|\varphi'(x_{p+j+1})\|} \right\| \leqslant \left\| \frac{\|\varphi'(x_{p+j+1})\|\,(\varphi'(x_{p+j}) - \varphi'(x_{p+j+1}))}{\|\varphi'(x_{p+j})\|\,\|\varphi'(x_{p+j+1})\|} + \frac{(\|\varphi'(x_{p+j+1})\| - \|\varphi'(x_{p+j})\|)\,\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j})\|\,\|\varphi'(x_{p+j+1})\|} \right\|,$$

whence, in view of the condition $\varphi(x) \in C^{1,1}(E_n)$, it follows that

$$\left\| \frac{\varphi'(x_{p+j})}{\|\varphi'(x_{p+j})\|} - \frac{\varphi'(x_{p+j+1})}{\|\varphi'(x_{p+j+1})\|} \right\| \leqslant \frac{L\langle\varphi'(x_{p+j}), s_{p+j}\rangle}{\|\varphi'(x_{p+j})\|\,L} + \frac{L\langle\varphi'(x_{p+j}), s_{p+j}\rangle}{\|\varphi'(x_{p+j})\|\,L} \leqslant 2\frac{\langle\varphi'(x_{p+j}), s_{p+j}\rangle}{\|\varphi'(x_{p+j})\|}.$$

Assume that the theorem is false, i.e.

$$2\langle\varphi'(x_{p+j}), s_{p+j}\rangle/\|\varphi'(x_{p+j})\| < 1/(2n\sqrt{n})$$

for all $j \in \{0, 1, \ldots, n-1\}$. Then, coming back to (3), we derive

$$\left\| \frac{\varphi'(\mathbf{x}_p)}{\|\varphi'(\mathbf{x}_p)\|} - \frac{\varphi'(\mathbf{x}_{p+i})}{\|\varphi'(\mathbf{x}_{p+i})\|} \right\| \leqslant n \frac{1}{2n\sqrt{n}} = \frac{1}{2\sqrt{n}}.$$

Therefore

$$\varphi'(\mathbf{x}_{p+i})/\|\varphi'(\mathbf{x}_{p+i})\| \in U_\delta(\varphi'(\mathbf{x}_p)/\|\varphi'(\mathbf{x}_p)\|),$$

$$\delta = 1/(2\sqrt{n}), \quad p \in I = \{0, 1, \ldots\},$$

for all $i \in \{0, \ldots, n-1\}$, and, consequently,

$$\varphi'(\mathbf{x}_{p+i_0})/\|\varphi'(\mathbf{x}_{p+i_0})\| \in U_\delta(\varphi'(\mathbf{x}_p)/\|\varphi'(\mathbf{x}_p)\|),$$

which, according to the corollary, means that the inequality

$$\langle \varphi'(\mathbf{x}_{p+i_0})/\|\varphi'(\mathbf{x}_{p+i_0})\|, \ \mathbf{s}_{p+i_0} \rangle \geqslant 1/(2\sqrt{n}) > 1/(4n\sqrt{n})$$

is fulfilled. We have arrived at a contradiction to the assumption that the inequality (2) is not fulfilled. Δ

Theorem 2. *Let the set* $X_0 = \{\mathbf{x}: \varphi(\mathbf{x}) \leqslant \varphi(\mathbf{x}_0)\}$ *be bounded, let* $\varphi(\mathbf{x}) \in C^{1,1}(E_n)$, *and let the sequence* $\{\mathbf{x}_k\}$ *be constructed by the formulas* (9.7.9), (9.13). *Then the following estimate for the rate of convergence takes place:*

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \delta n^4/m, \quad m \in I,$$

where $\delta > 0$ *is a constant not depending on the dimension of the space* E_n.

Proof. It follows from Theorem 1 that for any $p \in I$ there is at least one element in the sequence α^2_{p+i} $(i = 0, 1, \ldots, n-1)$

$$\alpha^2_{p+i} = \langle \varphi'(\mathbf{x}_{p+i}), \ \mathbf{s}_{p+i} \rangle^2 / \|\varphi'(\mathbf{x}_{p+i})\|^2,$$

such that $\alpha^2_{p+i} \geqslant 1/(16n^3)$. Therefore

$$\varphi(\mathbf{x}_{p+i}) - \varphi(\mathbf{x}_{p+i+1}) \geqslant \frac{1}{L} \alpha^2_{p+i} \|\varphi'(\mathbf{x}_{p+i})\|^2$$

$$\geqslant \frac{1}{16} \frac{1}{L} \frac{1}{n^3} \|\varphi'(\mathbf{x}_{p+i})\|^2,$$

and, consequently,

$$\frac{\varphi(\mathbf{x}_{p+i}) - \varphi(\mathbf{x}_{p+i+1})}{\|\varphi'(\mathbf{x}_{p+i})\|^2} \geqslant \frac{1}{16Ln^3}.$$

Thus, Theorem 9.3.3 implies

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \eta \left[\sum_{i=0}^{[m/n]} \frac{1}{16Ln^3} \right]^{-1} \leqslant \eta \frac{Ln^3 16}{m} = \frac{\delta n^4}{m}.$$

It remains to note that, according to Supplement 1, the parameter β_k is written $\beta_k = \frac{C_k}{L} \langle \varphi'(\mathbf{x}_k), \mathbf{s}_k \rangle$ where $L2^{-|N-1|} \leqslant C_k \leqslant L2^{|N-1|}$, which guarantees that the process of determining β_k by

the duplication method is finite. If β_k is replaced everywhere above by $\tilde{\beta}_k = \frac{C_k}{L} \langle \varphi'(x_k), s_k \rangle$, the assertions of Theorems 1 and 2 remain true. $\triangle$

If the function $\varphi(\mathbf{x})$ is strongly convex and the coefficient β_k is found from the condition

$$\beta_k = \arg\min_{\beta \geqslant 0} \varphi(\mathbf{x}_k - \beta s_k),$$

then the relation

$$\langle \varphi'(x_k), s_k \rangle / L \leqslant \beta_k \leqslant \langle \varphi'(x_k), s_k \rangle / \rho$$

holds, and therefore the foregoing results are valid for this way of selecting the steplength as well. In this case the estimate takes the form

$$\varphi(\mathbf{x}_m) - \varphi^* \leqslant \exp\left[-\rho\delta_1 \frac{m}{n^4}\right], \quad m = 0, 1, \ldots,$$

$$\| \mathbf{x}_m - \mathbf{x}^* \|^2 \leqslant \frac{2}{\rho} \exp\left[-\rho\delta_2 \frac{m}{n^4}\right], \quad m = 0, 1, \ldots,$$

where $\delta_1, \delta_2 > 0$ are some constants not depending on n and m.

Supplement 3. Degeneracy in Extremal Problems

The material in Chap. 3 was presented under the assumption that the constraints $f_i(\mathbf{x})$ $(i = \overline{1, m})$ and the objective function $\varphi(\mathbf{x})$ were not degenerate, i.e. $\varphi'(\mathbf{x}^*) \not\equiv 0$, $f_i'(\mathbf{x}^*) \not\equiv 0$ $(i = \overline{1, m})$. In the case of degeneracy the Kuhn-Tucker theorem becomes meaningless since

$$f_i'(\mathbf{x}^*) = 0, \quad i = \overline{1, m}, \tag{1}$$

$$\varphi'(\mathbf{x}^*) = 0. \tag{2}$$

In this supplement we present optimality criteria for degenerate mathematical programming problems. To simplify the presentation, we consider only the case of active constraints, i.e. $f_i(\mathbf{x}) = 0$ $(i = \overline{1, m})$. As before, we put $f^T(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))$ and formalize the degenerate problem in question in the following way:

$$\varphi(\mathbf{x}) \to \min_X, \quad X = \{\mathbf{x} \in E_n : \mathbf{f}(\mathbf{x}) = 0\} \tag{3}$$

under the conditions

$$\begin{aligned} &\varphi^{(q)}(\mathbf{x}^*) = 0, \quad q = \overline{1, p-1}, \\ &f_i^{(q)}(\mathbf{x}^*) = 0, \quad q = \overline{1, p-1}, \quad i = \overline{1, m}. \end{aligned} \tag{4}$$

We denote by $\varphi^{(p)}(\mathbf{x}_0)$ the pth derivative of the function $\varphi(\mathbf{x})$ at the point $\mathbf{x}_0$. For instance, $\varphi'(\mathbf{x}_0)$ is the gradient of $\varphi(\mathbf{x})$ at $\mathbf{x}_0$, $\varphi^{(2)}(\mathbf{x}) = \varphi''(\mathbf{x}_0)$ is the matrix of second derivatives, etc.

The expression $\varphi^{(p)}(\mathbf{x}_0)$ symbolizes an operator on the product space $\underbrace{E_n \times \ldots \times E_n}_{p}$ whose action is defined in the following way. Let $\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(p)}$ be elements of E_n. Then $\varphi^{(p)}(\mathbf{x}_0)[\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(p)}]$

$$\triangleq \left(\frac{\partial}{\partial x_1} y_1^{(1)} + \ldots + \frac{\partial}{\partial x_n} y_n^{(1)}\right) \ldots \left(\frac{\partial}{\partial x_1} y_1^{(p)} + \ldots + \frac{\partial}{\partial x_n} y_n^{(p)}\right) \varphi(x_0)$$

$$\triangleq \sum_{i_1 + \ldots + i_n = p} \frac{p!}{i_1! \ldots i_n!} \frac{\partial \varphi^{(p)}(\mathbf{x}_0)}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}} y_1^{(1)} \ldots y_{i_1}^{(1)} \ldots y_1^{(p)} \ldots y_{i_n}^{(p)}. \tag{5}$$

For example, if $\mathbf{y}^{(1)} = \ldots = \mathbf{y}^{(p)} = \mathbf{h}$, then,

$$\varphi^{(p)}(\mathbf{x}_0) \underbrace{\mathbf{h}, \ldots, \mathbf{h}}_{p} \triangleq \varphi^{(p)}(\mathbf{x}_0)[\mathbf{h}]^p$$

$$\triangleq \sum_{i_1 + \ldots + i_n = p} \frac{p!}{i_1! \ldots i_n!} \frac{\partial \varphi^{(p)}(\mathbf{x}_0)}{\partial x_1^{i_1} \ldots \partial x_n^{i_n}} h_1^{i_1} \ldots h_n^{i_n}. \tag{6}$$

We introduce the following notation:

$$\operatorname{Im} f^{(p)}(\mathbf{x}_0) = \{\mathbf{y} \in E_m : f_i^{(p)}(\mathbf{x}_0)[\mathbf{x}]^p = y_i, \quad i = \overline{1, m}, \quad \mathbf{x} \in E_n\},$$

$$\operatorname{Ker} f^{(p)}(\mathbf{x}_0) = \{\mathbf{x} \in E_n : f_i^{(p)}(\mathbf{x}_0)[\mathbf{x}]^p = 0, \quad i = \overline{1, m}\}.$$

I. The condition of p-regularity of constraints.

Definition. We shall say that a mapping $\mathbf{f}: E_n \to E_m$ is *p-regular at a point* $\mathbf{x}_0$ if

$$f^{(p)}(\mathbf{x}_0)[\mathbf{h}]^{p-1} E_n = E_m$$

for any $\mathbf{h} \in \operatorname{Ker} \mathbf{f}(\mathbf{x}_0)$ and $\mathbf{h} \neq 0$ (here the expression $f^{(p)}(\mathbf{x}_0)[\mathbf{h}]^{p-1} E_n$ means that the operator $f^{(p)}(\mathbf{x}_0)[\mathbf{h}]^{p-1}$ is applied to all the elements $\mathbf{z}$ of the space E_n).

In what follows we assume, without loss of generality, that $\mathbf{x}_0 = 0$, $f_i(\mathbf{x}_0) = 0$ $(i = \overline{1, m})$.

Let M be a subset in E_n. An element $\mathbf{x} \in E_n$ is called a tangent vector to the set M at a point $\mathbf{x}_0 \in M$ if there is $\varepsilon > 0$ and a mapping $t \to \mathbf{r}(t)$ of the closed interval $[0, \varepsilon]$ into E_n such that $\mathbf{x}_0 + t\mathbf{x} + \mathbf{r}(t) \in M$ for all $t \in [0, \varepsilon]$, $\|\mathbf{r}(t)\|/t \to 0$, $t \to 0$. The set of tangent vectors to the set M at $\mathbf{x}_0$ is called the tangent cone to M at the point $\mathbf{x}_0$ and is denoted $TM(\mathbf{x}_0)$.

II. Basic theorems. To prove the optimality conditions we shall need some classical theorems of functional analysis.

Theorem (the generalized principle of contracting mappings). *Let $\Phi: U(\mathbf{z}_0, r_1) \to E_n$ be a mapping of a ball $U(\mathbf{z}_0, r_1) = \{\mathbf{z} \mid \rho(\mathbf{z}, \mathbf{z}_0) < r_1\}$ $(r_1 > 0)$ into E_n, the set $\Phi(\mathbf{z})$ being nonempty for any $\mathbf{z} \in U(\mathbf{z}_0, r_1)$. Assume that there is θ $(0 < \theta < 1)$ such that*

(a) $h(\Phi(\mathbf{z}_1), \Phi(\mathbf{z}_2)) \leqslant \theta \rho(\mathbf{z}_1, \mathbf{z}_2)$ *for any* $\mathbf{z}_1, \mathbf{z}_2 \in U(\mathbf{z}_0, r_1)$,

(b) $\rho(\mathbf{z}_0, \Phi(\mathbf{z}_0)) < (1 - \theta) r_1$.

Then for any number r_2 satisfying the inequality $\rho(z_0, \Phi(z_0)) < r_2 < (1-0)\, r_1$ there is an element $z \in U(z_0, r_2/(1-\theta))$ such that $z \in \Phi(z)$. Moreover, among the points z satisfying these conditions there is one for which

$$\rho(z, z_0) \leqslant \frac{2}{1-\theta}\, \rho(z_0, \Phi(z_0)).$$

(Here $h(A, B) = \max\{\sup_{z \in A} \rho(z, B), \sup_{z \in B} \rho(z, A)\}$ is Housdorff's distance between the sets A and B.)

The open mapping theorem of Banach. *Let Λ be a linear operator on $E_n \to E_m$. Put*

$$C(\Lambda) = \sup_{y \in E_m} (\| y \|^{-1} \inf\{\| x \| \mid x \in E_n, \quad \Lambda x = y\}).$$

If $\operatorname{Im} \Lambda = E_m$, *then* $C(\Lambda) < \infty$.

The mean value theorem. *Let U be an open set in E_n and let a mapping $F\colon U \to E_m$ be differentiable at each point of the segment $[x, x+h] \subset U$. Then*

$$\| F(x+h) - F(x) \| \leqslant \sup_{0 \leqslant t \leqslant 1} \| F'(x+th) \| \| h \|.$$

III. The generalized Lyusternik theorem.

Theorem 1 (the generalized Lyusternik theorem). *Let U be a neighborhood of the point 0 and let f be a mapping of the set U into E_m differentiable up to order p inclusive and satisfying the condition (4). Assume that the mapping f is p-regular at 0, and let its pth derivative be continuous. Then the tangent cone to the set $M(0) = \{x \in U \mid f(x) = 0\}$ at 0 coincides with the kernel of the operator $f^{(p)}(0)$:*

$$TM(0) = \operatorname{Ker} f^{(p)}(0). \tag{7}$$

Moreover, if $\sup_{\|h\|=1} \| \{f^{(p)}(0)[h]^{p-1}\}^{-1} \| \leqslant k$, then there exist a neighborhood $U' \subset U$ of the point 0, a number $\delta > 0$, and a mapping $y \to x(y)$ of U' into E_n such that $f(y + x(y)) = f(0)$ and $\| x(y) \| \leqslant \delta \| f(y) - f(0) \|^{1/p}$ $\forall\, y \in U'$.

Proof. Let a vector x belong to $TM(0)$.

We shall prove that $x \in \operatorname{Ker} f^{(p)}(0)$. Indeed, $tx + r(t) \in M(0)$, $\| r(t) \|/t \to 0$ for $t \to 0$, i.e.

$$f(tx + r(t)) = f(0). \tag{8}$$

On the other hand,

$$f(tx + r(t)) = f(0) + \ldots + \frac{f^{(p)}(0)}{p!}[tx + r(t)]^p + \omega(t), \tag{9}$$

$$\| \omega(t) \|/t^p \to 0.$$

Equating the right-hand sides of (8) and (9) we find

$$f(0) = f(0) + \frac{f^{(p)}(0)[tx + r(t)]^p}{p!} + \omega(t),$$

whence

$$\frac{t^p}{p!} f^{(p)}(0)\left[\mathbf{x}+\frac{\mathbf{r}(t)}{t}\right]^p+\omega(t)=0. \tag{10}$$

Cancelling by t^p in (10) we obtain

$$\frac{1}{p!} f^{(p)}(0)\left[\mathbf{x}+\frac{\mathbf{r}(t)}{t}\right]^p+\frac{\omega(t)}{t^p}=0. \tag{11}$$

However, the left-hand side of (11) is equal to

$$f^{(p)}(0)[\mathbf{x}]^p + C_p^1 f^{(p)}(0)\left[\frac{\mathbf{r}(t)}{t}, \overbrace{\mathbf{x}, \ldots, \mathbf{x}}^{p-1}\right] + \ldots + f^{(p)}(0)\left[\frac{\mathbf{r}(t)}{t}\right]^p+\frac{\omega(t)}{t^p} p!=0. \tag{12}$$

Taking into consideration the continuity of $f^{(p)}(0)$ and the condition (4) we conclude from (12) for $t \to +0$ that $f^{(p)}(0)[\mathbf{x}]^p = 0$, i.e. $\mathbf{x} \in \operatorname{Ker} f^{(p)}(0)$, and thus $TM(0) \subset \operatorname{Ker} f^{(p)}(0)$.

Now, let $\mathbf{h} \in \operatorname{Ker} f^{(p)}(0)$, i.e. $f^{(p)}(0)[\mathbf{h}]^p = 0$. We assume, without loss of generality, that $\|\mathbf{h}\| = 1$. Put $\mathbf{x}_\alpha = \alpha\mathbf{h}$, $\alpha > 0$, and denote by $U_d(\mathbf{x}) = \{\mathbf{y} \in E_n \mid \|\mathbf{x}-\mathbf{y}\| \leqslant d\}$ a d-neighborhood of the point $\mathbf{x}$. We now consider the many-valued mapping

$$\Phi_{\mathbf{x}_\alpha}(\mathbf{x}) = \mathbf{x} - (p-1)!\{f^{(p)}(0)[\mathbf{x}_\alpha]^{p-1}\}^{-1}\mathbf{f}(\mathbf{x}_\alpha+\mathbf{x}),$$

where $f^{(p)}(0)[\mathbf{x}_\alpha]^{p-1}$ is a continuous linear operator for each fixed $\alpha > 0$.

We shall prove that for each sufficiently small $\alpha > 0$ the operator $\Phi_{\mathbf{x}_\alpha}(\mathbf{x})$ satisfies the following relations:

$$h(\Phi_{\mathbf{x}_\alpha}(\mathbf{x}_1), \Phi_{\mathbf{x}_\alpha}(\mathbf{x}_2)) \leqslant \gamma\|\mathbf{x}_1-\mathbf{x}_2\|, \quad \mathbf{x}_1, \mathbf{x}_2 \in U_{d(\alpha)}(0), \quad \gamma \in (0, 1); \tag{13a}$$

$$\rho(0, \Phi_{\mathbf{x}_\alpha}(0)) < d(\alpha)(1-\gamma). \tag{13b}$$

This will imply that the operator $\Phi_{\mathbf{x}_\alpha}(\mathbf{x})$ satisfies the conditions of the generalized contracting mapping principle and hence there exists a point $\mathbf{z}(\alpha)$ such that

$$\mathbf{z}(\alpha) \in \Phi_{\mathbf{x}_\alpha}(\mathbf{z}(\alpha)), \ \rho(\mathbf{z}(\alpha), 0) \leqslant 2\rho(0, \Phi_{\mathbf{x}_\alpha}(0))/(1-\gamma).$$

We first prove that

$$\rho(0, \Phi_{\mathbf{x}_\alpha}(0)) = o(\alpha) \tag{14}$$

for sufficiently small $\alpha > 0$. According to the open mapping theorem of Banach, we have

$$m = \sup_{\mathbf{y} \in E_m}(\|\mathbf{y}\|^{-1} \inf\{\|\mathbf{x}\| \mid \mathbf{x} \in E_n, \quad f^{(p)}(0)[\mathbf{h}]^{p-1}\mathbf{x} = \mathbf{y}\}) < \infty,$$

and therefore

$$m_\alpha = \sup_{y \in E_m} (\| y \|^{-1} \inf \{ \| x \| \mid x \in E_n,$$

$$f^{(p)}(0)[\alpha h]^{p-1} x = y\}) = \frac{m}{\alpha^{p-1}},$$

where $f^{(p)}(0)[\alpha h]^{p-1} x = \alpha^{p-1} f^{(p)}(0)[h]^{p-1} x$. Since $h \in \operatorname{Ker} f^{(p)}(0)$, we have

$$\rho(0, \Phi_{x_\alpha}(0)) = \| (p-1)! \{f^{(p)}(0)[\alpha h]^{p-1}\}^{-1} f(\alpha h) \|$$

$$\leqslant (p-1)!\, m \| f(\alpha h) \| \alpha^{1-p} = (p-1)!\, m \alpha^{1-p} \left\| f(0) \right.$$

$$\left. + f'(0)[\alpha h] + \ldots + \frac{1}{(p-1)!} f^{(p)}(0)[\alpha h]^p + \omega(\alpha h) \right\|$$

$$= (p-1)!\, \alpha^{1-p} m \| \omega(\alpha h) \|,$$

where $\| \omega(\alpha h) \| = o(\alpha^p)$. Therefore

$$\rho(0, \Phi_{x_\alpha}(0)) \leqslant (p-1)!\, \alpha^{1-p} \| \omega(\alpha h) \| = (p-1)!\, m\, o(\alpha), \qquad (15)$$

and the relation (14) is fulfilled.

We now prove the condition (13a), i.e. the fact that $\Phi_{x_\alpha}(x)$ is a contracting operator in the ball of radius $d(\alpha)$ with center at the point 0 for any sufficiently small $\alpha > 0$ (moreover, below it will be shown that $d(\alpha) = \alpha/R$ where $R > 1$ is a constant, i.e. the same for all sufficiently small $\alpha > 0$). This together with (15) will imply the validity of the condition (13b). We have

$$h(\Phi_{\alpha h}(x_1), \Phi_{\alpha h}(x_2))$$

$$= \inf \{ \| z_1 - z_2 \| \mid z_i \in \Phi_{\alpha h}(x_i),\ i = 1, 2 \}$$

$$= \inf \left\{ \| z_1 - z_2 \| \left| \frac{1}{(p-1)!} f^{(p)}(0)[\alpha h]^{p-1} z_i \right. \right.$$

$$\left. \left. = \frac{1}{(p-1)} f^{(p)}(0)[\alpha h]^{p-1} x_i - f(\alpha h + x_i),\quad i = 1, 2 \right\} \right.$$

$$\leqslant \frac{m}{\alpha^{p-1}} \left\| f(\alpha h + x_1) - f(\alpha h + x_2) \right.$$

$$\left. - \frac{1}{(p-1)} f^{(p)}(0)[\alpha h]^{p-1} (x_2 - x_1) \right\|.$$

By the mean value theorem,

$$\| f(\alpha h + x_1) - f(\alpha h + x_2) \|$$

$$\leqslant \sup_{\theta \in [0, 1]} \| f'(\alpha h + x_1 + \theta(x_2 - x_1)) \| \| x_2 - x_1 \|.$$

Therefore

$$\left\| f(\alpha h + x_1) - f(\alpha h + x_2) - \frac{1}{(p-1)!} f^{(p)}(0)\,[\alpha h]^{p-1}(x_2 - x_1) \right\|$$

$$\leqslant \sup_{\theta \in [0,1]} \left\| f'(\alpha h + x_1 + \theta(x_2 - x_1)) - \frac{1}{(p-1)!} f^{(p)}(0)\,[\alpha h]^{p-1} \right\| \|x_2 - x_1\|. \quad (16)$$

We assume that $\| x_1 \| \leqslant \alpha/R$, $\| x_2 \| \leqslant \alpha/R$, $R > 1$, where the constant R is chosen so that the condition (13a) holds for all sufficiently small $\alpha > 0$ simultaneously. Below it will be shown that R can in fact be chosen in this way, and the value of R will be indicated. Expanding $f'(\alpha h + x_1 + \theta(x_2 - x_1))$ at zero by Taylor's formula we obtain

$$f'(\alpha h + x_1 + \theta(x_2 - x_1)) = f'(0) + f^{(2)}(0)\,[\alpha h + x_1 + \theta(x_2 - x_1)] + \cdots + \frac{1}{(p-1)!} f^{(p)}(0)\,[\alpha h + x_1 + \theta(x_2 - x_1)]^{p-1} + \varepsilon(\alpha), \quad (17)$$

where $\| \varepsilon(\alpha) \| = o(\alpha^{p-1})$ since $\| \alpha h + x_1 + \theta(x^2 - x_1) \| \leqslant 4\alpha$. The first $p - 1$ terms in (17) are equal to zero, and therefore

$$f'(\alpha h + x_1 + \theta(x_2 - x_1)) = \frac{1}{(p-1)!} f^{(p)}(0)\,[\alpha h + x_1 + \theta(x_2 - x_1)]^{p-1} + \varepsilon(\alpha) = \frac{1}{(p-1)!} f^{(p)}(0)\,[\alpha h]^{p-1} + \frac{1}{(p-1)!} \sum_{i=1}^{p-1} C_{p-1}^{i} f^{(p)}(0)\,[\alpha h]^{p-1-i}\,[x_1 + \theta(x_2 - x_1)]^{i} + \varepsilon(\alpha). \quad (18)$$

We now estimate the value of $R > 1$ ($\| x_1 \| \leqslant \alpha/R$, $\| x_2 \| \leqslant \alpha/R$) for which the norm of the second term on the right-hand side of (18) is less than $\alpha^{p-1}/(4m)$. This can be done because

$$\left\| \frac{1}{(p-1)!} \sum_{i=1}^{p-1} C_{p-1}^{i} f^{(p)}(0)\,[\alpha h]^{p-1-i}\,[x_1 + \theta(x_2 - x_1)]^{i} \right\| \leqslant \frac{1}{(p-1)!} \| f^{(p)}(0) \| \sum_{i=1}^{p-1} C_{p-1}^{i} \alpha^{p-1-i} \| x_1 + \theta(x_2 - x_1) \|^{i} \leqslant \frac{4}{(p-1)!} \| f^{(p)}(0) \| \sum_{i=1}^{p-1} \frac{C_{p-1}^{i} \alpha^{p-1-i}}{R^{i}} \leqslant \frac{1}{(p-1)!} 4\alpha^{p-1} \| f^{(p)}(0) \| \sum_{i=1}^{p-1} \frac{C_{p-1}^{i}}{R}.$$

Putting $R \geqslant 2^{p+3} \|f^{(p)}(0)\| m/(p-1)!$ we obtain

$$\left\| \frac{1}{(p-1)!} \sum_{i=1}^{p-1} C_{p-1}^{i} f^{p}(0) [\alpha \mathbf{h}]^{p-1-i} [\mathbf{x}_1 + \theta(\mathbf{x}_2 - \mathbf{x}_1)]^{i} \right\| \leqslant \frac{\alpha^{p-1}}{4m}. \tag{19}$$

Based on (18) and (19), we estimate the right-hand side using the formula (16):

$$\sup_{\theta \in [0,1]} \|f'(\alpha \mathbf{h} + \mathbf{x}_1 + \theta(\mathbf{x}_2 - \mathbf{x}_1)) - \frac{1}{(p-1)!} f^{(p)}(0) [\alpha \mathbf{h}]^{p-1}\| \|\mathbf{x}_2 - \mathbf{x}_1\|$$

$$\leqslant \sup_{\theta \in [0,1]} \left\| \frac{1}{(p-1)!} \sum_{p=1}^{p-1} C_{p-1}^{i} f^{(p)}(0) [\alpha \mathbf{h}]^{p-1-i} [\mathbf{x}_1 + \theta(\mathbf{x}_2 - \mathbf{x}_1)]^{i} + \varepsilon(\alpha) \right\| \|\mathbf{x}_2 - \mathbf{x}_1\| \leqslant \frac{1}{4} \frac{\alpha^{p-1}}{m} \|\mathbf{x}_2 - \mathbf{x}_1\| + \|\varepsilon(\alpha)\| \|\mathbf{x}_2 - \mathbf{x}_1\|$$

$$\leqslant \frac{1}{2} \frac{\alpha^{p-1} \|\mathbf{x}_2 - \mathbf{x}_1\|}{m} \tag{20}$$

for all sufficiently small $\alpha > 0$. This implies that

$$h(\Phi_{\alpha \mathbf{h}}(\mathbf{x}_1), \Phi_{\alpha \mathbf{h}}(\mathbf{x}_2)) \leqslant \frac{m}{\alpha^{p-1}} \left\| \mathbf{f}(\alpha \mathbf{h} + \mathbf{x}_1) - \mathbf{f}(\alpha \mathbf{h} + \mathbf{x}_2) - \frac{1}{(p-1)!} f^{(p)}(0) [\alpha \mathbf{h}]^{p-1} (\mathbf{x}_2 - \mathbf{x}_1) \right\| \leqslant \frac{1}{2} \|\mathbf{x}_2 - \mathbf{x}_1\|,$$

$$\mathbf{x}_1, \mathbf{x}_2 \in U_{d(\alpha)}(0),$$

i.e. the condition (13) is fulfilled with $\gamma = 1/2$ and

$$R = 2^{p+3} \|f^{(p)}(0)\| m/(p-1)!$$

Thus, the conditions (13) have been proved, and therefore for any sufficiently small $\alpha > 0$ there is a point

$$\mathbf{z}(\alpha) \in \Phi_{\mathbf{x}_\alpha}(\mathbf{z}(\alpha)) \tag{21}$$

such that

$$\rho(\mathbf{z}(\alpha), 0) = \|\mathbf{z}(\alpha)\| \leqslant 4\rho \quad \left(0, \Phi_{\mathbf{x}_\alpha}(0)\right) = o(\alpha), \tag{22}$$

and (21) implies that $\mathbf{f}(\alpha \mathbf{h} + \mathbf{z}(\alpha)) = 0$, i.e. $\alpha \mathbf{h} + \mathbf{z}(\alpha) \in M(0)$. This together with (19) implies that $\mathbf{h} \in TM(0)$, and hence we have proved that $\operatorname{Ker} f^{(p)}(0) \subset TM(0)$.

We now prove the concluding assertion of the theorem. From the open mapping theorem of Banach and the condition that the operator $\{f^{(p)}(0)[\mathbf{h}]^{p-1}\}^{-1}$ is bounded it follows that

$$\sup_{\|\mathbf{h}\|=1} \sup_{\mathbf{y} \in E_m} (\|\mathbf{y}\|^{-1} \inf \{\|\mathbf{x}\| \mid \mathbf{x} \in E_n, f^{(p)}(0)[\mathbf{h}]^{p-1}\mathbf{x} = \mathbf{y}\}) = M < \infty.$$

Fix an arbitrary $\mathbf{h}$, $\|\mathbf{h}\| = 1$. Then for any mapping $\Phi_{\alpha\mathbf{h}}(\mathbf{x})$ we have

$$\begin{aligned}
&\mathbf{h}(\Phi_{\alpha\mathbf{h}}(\mathbf{x}_1),\ \Phi_{\alpha\mathbf{h}}(\mathbf{x}_2)) \\
&\quad = \inf\{\|\mathbf{z}_1 - \mathbf{z}_2\| \mid \mathbf{z}_i \in \Phi_{\alpha\mathbf{h}}(\mathbf{x}_i),\ i = 1,\ 2,\} \\
&\quad = \inf\{\|\mathbf{z}_1 - \mathbf{z}_2\| \mid f^{(p)}(0)[\alpha\mathbf{h}]^{p-1}\mathbf{z}_i \\
&\quad = f^{(p)}(0)[\alpha\mathbf{h}]^{p-1}\mathbf{x}_i - \mathbf{f},\ (\alpha\mathbf{h} + \mathbf{x}_i)\ i = 1,\ 2\} \\
&\quad \leqslant \frac{M}{\alpha^{p-1}} \Big\| \mathbf{f}(\alpha\mathbf{h} + \mathbf{x}_1) - \mathbf{f}(\alpha\mathbf{h} + \mathbf{x}_2) \\
&\qquad - \frac{1}{(p-1)!} f^{(p)}(0)[\alpha\mathbf{h}]^{p-1}(\mathbf{x}_2 - \mathbf{x}_1) \Big\|.
\end{aligned}$$

This inequality automatically implies, by analogy with (61)-(20) that for all $\mathbf{x}_1$, $\mathbf{x}_2$ belonging to a neighborhood u' of zero having a sufficiently small radius r and any $\mathbf{y} \in u'$ there holds

$$h(\Phi_{\mathbf{y}}(\mathbf{x}_1),\ \Phi_{\mathbf{y}}(\mathbf{x}_2)) \leqslant \|\mathbf{x}_1 - \mathbf{x}_2\|/2. \tag{23}$$

Further, by virtue of (15),

$$\rho(0,\ \Phi_{\mathbf{y}}(0)) \leqslant M\|\mathbf{f}(\mathbf{y})\|^{1/p} \leqslant r/2. \tag{24}$$

From (23) and (24) it follows that for any $\mathbf{y} \in u'$ the mappings $\Phi_{\mathbf{y}}(\mathbf{x})$ satisfy the conditions of the generalized contracting mapping principle. Therefore there is a vector $\mathbf{x} = \mathbf{x}(\mathbf{y})$ such that $\mathbf{x}(\mathbf{y}) \in \Phi_{\mathbf{y}}(\mathbf{x}(\mathbf{y}))$, and, consequently, $\mathbf{f}(\mathbf{y} + \mathbf{x}(\mathbf{y})) = 0$. On the other hand,

$$\|\mathbf{x}(\mathbf{y})\| \leqslant \delta_1 \rho(0,\ \Phi_{\mathbf{y}}(0)) \leqslant \delta\|\mathbf{f}(\mathbf{y})\|^{1/p},$$

where δ_1, δ_2 are positive constants not depending on $\mathbf{y}$. The theorem is proved. △

IV. *p*th-order necessary and sufficient conditions for optimality.

Theorem 2. *Let U be an open set in E_n and let a function φ: $U \to R$ and a mapping $\mathbf{f}$: $E_n \to E_m$ be strictly differentiable up to order p inclusive at a point $\mathbf{x}^* \in U$. If $\mathbf{x}^*$ yields local minimum in the problem* (3)-(4) *and if $\mathbf{f}$ is a regular mapping of the pth order at the point $\mathbf{x}^*$, then*

$$\varphi^{(p)}(\mathbf{x}^*)[\mathbf{h}]^p \geqslant 0 \quad \forall h \in \operatorname{Ker} f^{(p)}(\mathbf{x}^*). \tag{25}$$

In case $\sup_{\|\mathbf{h}\|=1} \|\{f^{(p)}(\mathbf{x}^*)[\mathbf{h}]^{p-1}\}^{-1}\| \leqslant k$ *and the inequality*

$$\varphi^{(p)}(\mathbf{x}^*)[\mathbf{h}]^p \geqslant \alpha\|\mathbf{h}\|^p, \quad \mathbf{h} \in \operatorname{Ker} f^{(p)}(\mathbf{x}^*) \setminus \{0\}, \tag{26}$$

is fulfilled for some $\alpha > 0$, $\mathbf{x}^*$ is a point of local minimum.

Proof. Necessity. Let $\mathbf{x}^*$ be the solution to the problem (3) and let $\mathbf{h} \in \operatorname{Ker} f^{(p)}(\mathbf{x}^*)$. According to Theorem 1, $\mathbf{h} \in TM(\mathbf{x}^*)$ where $M(\mathbf{x}^*) = \{\mathbf{x} \in E_n \mid \mathbf{f}(\mathbf{x}) = 0\}$, i.e. there exists a mapping $\mathbf{r}(\cdot)$: $[-\varepsilon,\ \varepsilon] \to M(\mathbf{x}^*)$ such that

$$\mathbf{f}(\mathbf{x}^* + t\mathbf{h} + \mathbf{r}(t)) = 0, \quad \|\mathbf{r}(t)\|/t \to 0, \quad t \to 0.$$

This means that $\mathbf{x}^* + t\mathbf{h} + \mathbf{r}(t)$ is a feasible element in the problem for $t \in [-\varepsilon, \varepsilon]$ and, consequently,

$$\varphi(\mathbf{x}^*) \leqslant \varphi(\mathbf{x}^* + t\mathbf{h} + r(\mathbf{t})).$$

Therefore

$$\varphi(\mathbf{x}^*) \leqslant \varphi(\mathbf{x}^* + t\mathbf{h} + \mathbf{r}(t)) = \varphi(\mathbf{x}^*) + \varphi'(\mathbf{x}^*)[t\mathbf{h} + \mathbf{r}(t)] + \ldots + \frac{1}{p!}\varphi^{(p)}(\mathbf{x}^*)[t\mathbf{h} + \mathbf{r}(t)]^p + o(t^p) = \varphi(\mathbf{x}^*) + \frac{t^p}{p!}\varphi^{(p)}(\mathbf{x}^*)[\mathbf{h}]^p + o(t^p).$$

This immediately implies (25).

Sufficiency. Assume that $\varphi(\mathbf{x}^*) = 0$. Denote by $B(\mathbf{h}_1, \ldots, \mathbf{h}_p)$ the multilinear form $(p!)^{-1}\varphi^{(p)}(\mathbf{x}^*)[\mathbf{h}_1, \ldots, \mathbf{h}_p]$. It is obvious that the form $B(\mathbf{h}_1, \ldots, \mathbf{h}_p)$ is continuous and symmetric. We take $\varepsilon > 0$ such that

$$\gamma(\varepsilon) = \left(\frac{\alpha}{p!}(1-\varepsilon)^p - C_p^1\|B\|(1+\varepsilon)^{p-1}\varepsilon - \ldots - C_p^k\|B\|(1+\varepsilon)^{p-k}\varepsilon^k - \ldots - C_p^p\|B\|\varepsilon^p\right) - \frac{\alpha}{2p!} > 0.$$

(Since $\gamma(0) = \alpha/(2p!) > 0$, such a number $\varepsilon > 0$ is sure to exist.) By the hypothesis, the functions $\mathbf{f}(\mathbf{x})$ and $\varphi(\mathbf{x})$ possess the derivatives of the pth order with respect to $\mathbf{x}$ at the point $\mathbf{x}^*$. Using Taylor's formula and taking the condition (3) into account we see that $\varphi^i(\mathbf{x}^*) = 0$ $(i = \overline{1, p-1})$, and, consequently, there is $\delta > 0$ such that

$$|\varphi(\mathbf{x}^*) - B[\mathbf{h}]^p| < \frac{\alpha}{2p!}\|\mathbf{h}\|^p$$

for $\|\mathbf{h}\| \leqslant \delta$. Now, let $\|\mathbf{h}\| < \delta$ and let $\mathbf{x}^* + \mathbf{h}$ be a feasible element in the problem, i.e. $\mathbf{f}(\mathbf{x}^* + \mathbf{h}) = \mathbf{0}$. Then applying the generalized Lyusternik theorem to the functional $F(\mathbf{x}) = f^{(p)}(\mathbf{x}^*)[\mathbf{h} + \mathbf{x}]^p$ for $p = 1$ we prove the existence of an element $\tilde{\mathbf{h}}$ such that

$$f^{(p)}(\mathbf{x}^*)[\mathbf{h} + \tilde{\mathbf{h}}]^p = 0,$$
$$\|\tilde{\mathbf{h}}\| < k_1 \frac{\|f^{(p)}(\mathbf{x}^*)[\mathbf{h}]^p\|}{\|\mathbf{h}\|^{p-1}} = o(\|\mathbf{h}\|). \tag{27}$$

Hence, $\mathbf{h} + \tilde{\mathbf{h}} \in \operatorname{Ker} f^{(p)}(\mathbf{x}^*)$, and, as follows from (27), $\|\mathbf{h} - (\mathbf{h} + \tilde{\mathbf{h}})\| = \|\tilde{\mathbf{h}}\| \leqslant \varepsilon\|\mathbf{h}\|$. This implies $(1 - \varepsilon)\|\mathbf{h}\| \leqslant \|\mathbf{h} + \tilde{\mathbf{h}}\| \leqslant (1 + \varepsilon)\|\mathbf{h}\|$, and, finally, we obtain

$$\varphi(\mathbf{x}^* + \mathbf{h}) - \varphi(\mathbf{x}^*) \geqslant B[\mathbf{h}]^p - \frac{\alpha}{2p!}\|\mathbf{h}\|^p = B[\mathbf{h} + \tilde{\mathbf{h}} - \tilde{\mathbf{h}}] - \frac{\alpha}{2p!}\|\mathbf{h}\|^p \geqslant B[\mathbf{h} + \tilde{\mathbf{h}}]^p - C_p^1\|B\|\,\|\mathbf{h} + \tilde{\mathbf{h}}\|^{p-1}\|\tilde{\mathbf{h}}\|$$

$$- \ldots - \|B\| \|\tilde{\mathbf{h}}\|^p - \frac{\alpha}{2p!} \|\mathbf{h}\|^p$$

$$\geqslant \left(\frac{1}{p!} \left(\alpha (1-\varepsilon)^p - C_p^1 \|B\| (1+\varepsilon)^{p-1} \varepsilon - \right. \right.$$

$$\left. \left. \ldots - \|B\| \varepsilon^p - \frac{\alpha}{p!} \right) \|\mathbf{h}\|^p \right) > 0,$$

i.e. $\mathbf{x}^*$ is the solution to the problem (3)-(4). The theorem is proved. Δ

Supplement 4. On the Penalty Function Method (Computational Aspect)

Theorem 1. *If functions* $\varphi(\mathbf{x})$ *and* $\psi(\mathbf{x})$ *are convex throughout the space* E_n *and the set* $Y = \operatorname{Arg\,min} \{\varphi(\mathbf{x}): \mathbf{x} \in X\}$ *is nonempty and bounded, then there is a sufficiently small* $\beta_0 > 0$ *such that the sets* $Y_\beta = \operatorname{Arg\,min} \{M(\mathbf{x}, \beta): \mathbf{x} \in E_n\}$ *are bounded jointly for all* $\beta \in (0, \beta_0]$.

Proof. Consider a point $\mathbf{y} \in Y$ and embed the set Y in a ball H of finite radius r with center at the point $\mathbf{y}$ so that $Y \subset \operatorname{int} H$. Denote by G the boundary of the set H.

If we prove that $M(\mathbf{z}, \beta) > M(\mathbf{y}, \beta) = \varphi(\mathbf{y})$ for all $\mathbf{z} \in G$, $\beta \in (0, \beta_0]$, then the convexity of the function $M(\mathbf{x}, \beta)$ would imply the inequality $M(\mathbf{x}, \beta) > M(\mathbf{y}, \beta)$ for all $\mathbf{x} \notin H$. Indeed, if we assume that there is $\mathbf{x} \notin H$ such that $M(\mathbf{x}, \beta) \leqslant M(\mathbf{y}, \beta)$ then the segment $[\mathbf{x}, \mathbf{y}]$ contains a point $\mathbf{z} = \lambda \mathbf{x} + (1 - \lambda)\, \mathbf{y} \in G$ $(0 < \lambda < 1)$, and this leads to the inequality $M(\mathbf{z}, \beta) \leqslant \lambda M(\mathbf{x}, \beta) + (1-\lambda) M(\mathbf{y}, \beta) \leqslant M(\mathbf{y}, \beta)$ contradicting the condition $M(\mathbf{z}, \beta) > M(\mathbf{y}, \beta)$. The fact that $M(\mathbf{x}, \beta) > M(\mathbf{y}, \beta)$ $\forall\, \mathbf{x} \notin H$ implies the inclusion $Y_\beta \subset H$ since $M(\mathbf{y}_\beta, \beta) \leqslant M(\mathbf{y}, \beta)$ (see Sec. 6.2.2). The existence of $\mathbf{y}_\beta$ for all $\beta \in (0, \beta_0]$ is obvious since the function $M(\mathbf{x}, \beta)$ attains minimum at a point $\mathbf{y}_\beta$ on the closed and bounded set H, and outside this set there holds $M(\mathbf{x}, \beta) > \varphi(\mathbf{y}) = M(\mathbf{y}, \beta) \geqslant M(y_\beta, \beta)$. Thus, we must prove that $M(\mathbf{z}, \beta) > M(\mathbf{y}, \beta)$ $\forall\, \mathbf{z} \in G$, $\beta \in (0, \beta_0]$.

(1) Assume that the set X is bounded and the radius r of the sphere G is such that $X \subset \operatorname{int} H$. Denote $\mu = \min\limits_{\mathbf{z} \in G} \varphi(\mathbf{z})$ and $\nu = \min\limits_{\mathbf{z} \in G} \psi(\mathbf{z})$. Since $\nu > 0$, there is $\beta_0 > 0$ such that

$$\frac{1}{\beta} \psi(\mathbf{z}) \geqslant \frac{1}{\beta_0} \psi(\mathbf{z}) \geqslant \frac{1}{\beta_0} \nu > \varphi(\mathbf{y}) - \mu$$

for all $\mathbf{z} \in G$ and $\beta \in (0, \beta_0]$, and, consequently,

$$M(\mathbf{z}, \beta) = \varphi(\mathbf{z}) + \frac{1}{\beta} \psi(\mathbf{z}) > \mu + \varphi(\mathbf{y}) - \mu = M(\mathbf{y}, \beta).$$

(2) Let the set X be unbounded. The set Y being bounded, it follows that $\varphi(\mathbf{z}) \geqslant \varphi(\mathbf{y}) + \Delta$ $\forall\, \mathbf{z} \in G \cap X$ where $0 < \Delta \leqslant \min\limits_{\mathbf{z} \in G \cap X} \varphi(\mathbf{z}) - \varphi(\mathbf{y})$.

Denote

$$U_\delta(X)=\{x: \rho(x, X)\leqslant\delta\},\quad G_1=G\cap U_\delta(X),\quad G_2=G\setminus U_\delta(X),$$

then $G=G_1\cup G_2$. We take $\delta=\delta(\Delta)>0$ such that $\varphi(z)>\varphi(y)+\frac{1}{2}\Delta\ \forall\, z\in G_1$. This is possible since the convex function $\varphi(x)$ on E_n is continuous. Thus, for all $z\in G_1$ we have

$$M(z, \beta)=\varphi(z)+\frac{1}{\beta}\psi(z)\geqslant\varphi(y)+\frac{1}{2}\Delta+\frac{1}{\beta}\psi(z)>\varphi(y)=M(y, \beta).$$

Further, denoting $\mu=\min\limits_{z\in G_2}\varphi(z)$ and $\nu=\min\limits_{z\in G_2}\psi(z)>0$ and repeating the argument in (1) we obtain

$$M(z, \beta)>M(y, \beta)\quad \forall z\in G_2,\ \beta\in(0, \beta_0].\quad\Delta$$

Theorem 2. *If the functions $\varphi(x)$ and $\psi(x)$ are convex and differentiable throughout the space E_n and the set $Y=\operatorname{Arg}\min\{\varphi(x): x\in X\}$ is nonempty and bounded, then there is a sufficiently small $\beta_0>0$ such that*

$$\forall\varepsilon>0\quad \exists\beta_0=\beta_0(\varepsilon)>0,\quad \delta=\delta(\varepsilon)>0$$

$$\forall x\notin U_\varepsilon(Y),\quad \beta\in(0, \beta_0]:\ \|\nabla M(x, \beta)\|\geqslant\delta.$$

Proof. Since the sets Y_β are bounded jointly for $\beta\in(0, \beta_0]$, the assertion of the theorem follows from Theorem 1 and Sec. 9.2.7. Δ

Theorem 3. *If*

(1) *the functions $\varphi(x)$ and $\psi(x)$ are convex and differentiable in E_n,*

(2) *the set $Y=\operatorname{Arg}\min\{\varphi(x): x\in X\}$ is nonempty and bounded,*

(3) *the sequence $\{x_k=x_k(\beta)\}$ is such that $\|\nabla M(x_k, \beta)\|\leqslant\varepsilon_k$, then $\rho(x_k, Y)\to 0$ for $\varepsilon_k\to 0$, $k\to\infty$, and $\beta\to 0$.*

Proof. We first prove that the sequence $\{x_k\}$ is bounded. By Theorem 1, there exists $\beta_0>0$ such that the sets Y_β are bounded jointly for all $\beta\in(0, \beta_0]$. According to Theorem 2, if a ball H with boundary G of radius ε such that $Y_\beta\subset\operatorname{int}H$ for all $\beta\in(0, \beta_0]$ is taken as $U_\varepsilon(Y)$, then $\exists\delta=\delta(\varepsilon)>0\ \forall x\notin H: \|\nabla M(x, \beta)\|\geqslant\delta$. If we assume that $\{x_k\}$ is unbounded, then $\{x_k\}\to\infty$, $k\to\infty$, and, consequently, there is an index k_0 such that $x_k\notin H$ and $\|\nabla M(x_k, \beta)\|\geqslant\delta$ for all $k\geqslant k_0$, which contradicts the condition (3).

We now assume that the theorem is false, i.e. there are a number $\Delta>0$ and a sequence $\{\beta_k\}\to 0$, $k\to\infty$, such that the inequality

$$|\varphi(x_k)-\varphi(y)|\geqslant\Delta\quad\text{for any}\quad y\in Y$$

holds for all $k=0, 1, \ldots$.

Since $\psi(x)\geqslant 0$ for all x and $\psi(y)=0$, we see that, in view of the boundedness of the sequence $\{x_k\}$, there must be $\|x_k-y_{\beta_k}\|\leqslant\eta<\infty$, and, using the convexity of the function $M(x, \beta)$, we

derive

$$\varphi(x_k) \leqslant \varphi(x_k) + \frac{1}{\beta_k}\psi(x_k) = M(x_k, \beta_k)$$
$$\leqslant M(y_{\beta_k}, \beta_k) + \langle \nabla M(x_k, \beta_k), x_k - y_{\beta_k}\rangle$$
$$\leqslant M(y_{\beta_k}, \beta_k) + \eta\varepsilon_k \leqslant M(y, \beta_k) + \eta\varepsilon_k = \varphi(y) + \eta\varepsilon_k.$$

Since the sequence $\{x_k\}$ is bounded, we assume, without changing the indexing, that $\lim_{k\to\infty} x_k = \tilde{x}$. Therefore $\varphi(\tilde{x}) \leqslant \varphi(y)$. It follows from Sec. 6.2.4 that $\psi(\tilde{x}) = \lim_{k\to\infty} \psi(x_k) = 0$, i.e. $\tilde{x} \in Y$, which contradicts the assumption that $|\varphi(x_k) - \varphi(\tilde{x})| \geqslant \Delta$. $\triangle$

Theorem 4. *If the following conditions hold:*

(a) $\varphi(x)$ *is a convex and differentiable function on* E_n,

(b) $f_i(x)$ $(i = \overline{1, m})$ *are concave and differentiable functions on* E_n,

(c) *the set* $Y = \operatorname{Arg\,min}\{\varphi(x): x \in X\}$ *is nonempty and bounded*,

(d) $\psi(x) = \sum_{i=1}^{m} |\min\{f_i(x), 0\}|^p$, $p > 1$,

(e) $|\varphi(u) - \varphi(v)| \leqslant L\rho(u, v)$ *for any* u *and* v *belonging to* E_n,

(f) $\exists \delta > 0$, $\alpha > 0$ $\forall x \in U_\delta(X) \setminus X$: $|f(x)| \geqslant \alpha\rho(x, X)$ *where* $f(x) = \min_{i=\overline{1, m}} f_i(x)$, $U_\delta(X) = \{x: \rho(x, X) \leqslant \delta\}$,

(g) $\|\nabla M(x_k, \beta)\| \leqslant \varepsilon_k$ $(k = 0, 1, \ldots)$ *for sufficiently small* $\beta > 0$, *then*

$$|\varphi(y) - M(x_k, \beta)| \leqslant \gamma(\beta^{1/(p-1)} + \varepsilon_k), \qquad \gamma = \text{const} > 0.$$

Proof. It follows from Theorem 1 that for sufficiently small $\beta > 0$ there exist points

$$y_\beta \in Y_\beta = \operatorname{Arg\,min}\{M(x, \beta): x \in E_n\}.$$

Since $\varphi(y) - M(y_\beta, \beta) \geqslant 0$ (see Sec. 6.2.2) and $M(x_k, \beta) - M(y_\beta, \beta) \geqslant 0$, we have

$$|\varphi(y) - M(x_k, \beta)|$$
$$\leqslant |\varphi(y) - M(y_\beta, \beta)| + |M(y_\beta, \beta) - M(x_k, \beta)|$$
$$= (\varphi(y) - M(y_\beta, \beta)) + (M(x_k, \beta) - M(y_\beta, \beta)).$$

Since the conditions of Theorem 6.2.7 are fulfilled, the inequality

$$\varphi(y) - M(y_\beta, \beta) \leqslant C\beta^{1/(p-1)}$$

takes place, and the condition (g) implies

$$M(x_k, \beta) - M(y_\beta, \beta) \leqslant \langle \nabla M(x_k, \beta), x_k - y_\beta\rangle$$
$$\leqslant \|\nabla M(x_k, \beta)\| \|x_k - y_\beta\| \leqslant \eta\varepsilon_k,$$

whence it follows that the assertion of the theorem is true for $\gamma = \max\{C, \eta\}$. $\triangle$

The computational aspect. We shall construct the sequences $\{\varepsilon_k\}$, $\{\beta_k\}$, and $\{x_k\}$ in the following way.

Let the parameters ε_k and β_k and the point x_{k-1} be known. We note that, generally, ε_0, β_0, and x_0 can be taken arbitrarily although it is preferable to determine them with regard to the available information about the functions $\varphi(x)$ and $\psi(x)$. We construct the point x_k so that the condition $\| \nabla M(x_k, \beta_k) \| \leqslant \varepsilon_k$ holds. Then we choose $\varepsilon_{k+1} < \varepsilon_k$ and $\beta_{k+1} < \beta_k$ so that the fulfilment of the conditions $\varepsilon_k \to 0$ and $\beta_k \to 0$ for $k \to \infty$ is guaranteed.

To find the point x_k we construct a relaxation sequence $\{y_i(k)\}$ beginning with the point $y_0 = x_{k-1}$ until a point $y_j = y_j(k)$ is calculated such that $\| \nabla M(y_j, \beta_k \| \leqslant \varepsilon_k$, i.e. $j = \min\{i: \| M(y_i, \beta_k) \| \leqslant \varepsilon_k, i = 1, 2, \ldots\}$. After that we put $x_k = y_j$.

It should be noted that, when constructing the relaxation sequence $\{y_i\}$ by the gradient descent method (see Secs. 9.5.1 and 9.5.2), the expression $\nabla M(y_i, \beta_k)$ is calculated at each iteration, and therefore the estimation of $\| \nabla M(y_i, \beta_k) \|$ involves no additional cumbersome calculations.

The convergence of the indicated method was proved in Theorem 4.

The so-called *iterative penalty function method* finds rather extensive application. The basic idea of the method consists in making the parameters α_k and β_k compatible so that for each value of β_k only one step of the gradient method

$$x_{k+1} = x_k - \alpha_k M'(x_k, \beta_k), \quad \alpha_k > 0,$$

is required to make the sequence $\{x_k\}$ converge to the solution of the problem.

The following conditions guarantee the indicated convergence:

$\varphi(x)$ is a strongly convex and differentiable function on E_n,

$f_i(x)$ $(i = \overline{1, m})$ are concave and differentiable functions on E_n,

$$\psi(x) = \sum_{i=1}^{m} [\min\{f_i(x), 0\}]^2,$$

x_0 is an arbitrary point,

$$\lim_{k\to\infty} \alpha_k = 0; \quad \lim_{k\to\infty} \beta_k = 0; \quad \lim_{k\to\infty} \frac{1/\beta_{k+1} - 1/\beta_k}{\alpha_k} = 0; \quad \lim_{k\to\infty} \frac{\alpha_k}{\beta_k^2} = 0.$$

Supplement 5. Stabilizing Properties of the Gradient Method

Let the function $\varphi(x)$ be convex and differentiable and let $Y = \operatorname{Arg\,min}\{\varphi(x): x \in E_n\}$. Consider the sequence

$$x_{k+1} = x_k - \beta_k \varphi'(x_k), \quad k = 0, 1, \ldots, \tag{1}$$

where the parameter $\beta_k > 0$ is determined from the condition

$$\varphi(x_k) - \varphi(x_k - \beta_k \varphi'(x_k)) \geqslant \varepsilon \beta_k \| \varphi'(x_k) \|^2, \quad \varepsilon \in (1/2, 1). \tag{2}$$

The computational procedure of finding a scalar $\beta_k > 0$ satisfying (2) was described in Supplement 1 where it was also shown that the determination of β_k by the duplication method is performed after a finite number of steps, and in this case $\beta_k \geqslant \beta > 0$.

In this supplement we consider the case when the set Y is a linear manifold. We first prove that the choice of $\mathbf{x}_0$ guarantees the convergence of the approximations constructed by the gradient method to the normal, relative to $\mathbf{x}_0$, solution $\mathbf{p}_0$ of the original problem*.

Theorem 1. *Let Y be a linear manifold, $\mathbf{p}_0$ be the orthogonal projection of the point $\mathbf{x}_0$ on Y, and $\{\mathbf{x}_k\}$ be constructed according to the scheme* (1), (2). *Then*

$$\mathbf{p}_0 = \tilde{\mathbf{y}}(\mathbf{x}_0) = \tilde{\mathbf{y}}, \quad \tilde{\mathbf{y}} = \lim_{k\to\infty} \mathbf{x}_k.$$

Proof. The existence of the point $\tilde{\mathbf{y}}$ was proved in Sec. 9.5.1. Consider the point $\mathbf{v}(\alpha) = \tilde{\mathbf{y}} + \alpha(\mathbf{p}_0 - \tilde{\mathbf{y}}) \in Y$, $\alpha > 0$. Since $\| \mathbf{p}_0 - \tilde{\mathbf{y}} \|^2 = \langle \tilde{\mathbf{y}} - \mathbf{p}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle$ and $\| \mathbf{v}(\alpha - \tilde{\mathbf{y}}) \| \leqslant \| \mathbf{v}(\alpha) - \mathbf{x}_k \|$ for any $k = 0, 1, \ldots$, we have

$$\begin{aligned} \| \mathbf{v}(\alpha) - \tilde{\mathbf{y}} \|^2 &= \| \mathbf{v}(\alpha) - \mathbf{x}_0 \|^2 - 2\langle \mathbf{v}(\alpha) - \mathbf{x}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle + \| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2 \\ &> \| \mathbf{v}(\alpha) - \tilde{\mathbf{y}} \|^2 - 2\langle \mathbf{v}(\alpha) - \mathbf{x}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle + \| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2, \end{aligned}$$

i.e.

$$\begin{aligned} \| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2 < 2\langle \mathbf{v}(\alpha) - \mathbf{x}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle &= 2\langle \tilde{\mathbf{y}} - \mathbf{x}_0 + \alpha(\mathbf{p}_0 - \tilde{\mathbf{y}}), \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle \\ &= 2\| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2 - 2\alpha\langle \tilde{\mathbf{y}} - \mathbf{p}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle \end{aligned}$$

whence

$$\alpha\langle \tilde{\mathbf{y}} - \mathbf{p}_0, \tilde{\mathbf{y}} - \mathbf{x}_0 \rangle < \| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2/2.$$

It follows that $\| \mathbf{p}_0 - \tilde{\mathbf{y}} \|^2 < \| \tilde{\mathbf{y}} - \mathbf{x}_0 \|^2/(2\alpha)$ for any $\alpha > 0$, i.e. $\mathbf{p}_0 = \tilde{\mathbf{y}}(\mathbf{x}_0)$. $\triangle$

We shall consider a consistent linear algebraic system of equations $A\mathbf{x} = \mathbf{u}$, where A is an $m \times n$ matrix, $\mathbf{x} \in E_n$, and $\mathbf{u} \in E_n$, and its approximate representation $\tilde{A}\mathbf{x} = \tilde{\mathbf{u}}$ on condition that $\| A - \tilde{A} \| \leqslant \delta$ and $\| \mathbf{u} - \tilde{\mathbf{u}} \| \leqslant \delta$.

Put $\varphi(\mathbf{x}) = \| \tilde{A}\mathbf{x} - \tilde{\mathbf{u}} \|^2$.

The scheme of the method.

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \beta_k \tilde{\varphi}'(\mathbf{x}_k), \quad k = 0, 1, \ldots, \tag{1}$$

$$\tilde{\varphi}(\mathbf{x}_k) - \tilde{\varphi}(\mathbf{x}_k - \beta_k \tilde{\varphi}'(\mathbf{x}_k)) \geqslant \varepsilon\beta_k \| \tilde{\varphi}'(\mathbf{x}_k) \|^2, \quad \varepsilon \in (1/2, 1), \tag{2}$$

if $\tilde{\varphi}(\mathbf{x}_k) > \delta$, then the point $\mathbf{x}_{k+1}$ is calculated, (3)

if $\tilde{\varphi}(\mathbf{x}_k) \leqslant \delta$, then the minimization process is ended. (4)

* According to the notation in Chap. 7, $\mathbf{p}_0 \equiv \mathbf{y}^0$, $\mathbf{x}_0 \equiv \mathbf{x}^0$.

The indicated process can be written as follows: find a point $\tilde{x} \in \tilde{Y} = \operatorname{Arg\,min} (\tilde{\varphi}(x) - \delta)_+$ where z_+ denotes the positive part of the vector z, i.e. $(z_+)_i = \max\{z_i, 0\}$, $i = \overline{1, n}$.

Lemma 1. *There is* $\tilde{\delta} > 0$ *such that* $p_0 \in \tilde{Y}$ *for all* $\delta \in (0, \tilde{\delta}]$.

Proof. We put $\tilde{\delta} = (\| p_0 \| + 1)^{-2}$. Then for all $\delta \in (0, \tilde{\delta}]$ the inequalities

$$\| \tilde{A}p_0 - \tilde{u} \|^2 = \| \tilde{A}p_0 - Ap_0 + u - \tilde{u} \|^2$$
$$\leqslant (\| (\tilde{A} - A) p_0 \| + \| u - \tilde{u} \|)^2 \leqslant (\| \tilde{A} - A \| \, \| p_0 \| + \| u - \tilde{u} \|)^2$$
$$\leqslant \delta \| p_0 \| + \delta)^2 = \delta^2 (\| p_0 \| + 1)^2 \leqslant \delta$$

are fulfilled. $\triangle$

Let $\{e_1, \ldots, e_n\}$ be a basis in the space E_n and let $L(e_1, \ldots, e_{r-1})$ be the subspace generated by the vectors $e_1, \ldots, e_{r-1}$.

Lemma 2. *Let* $Ap_0 = u$, $\| A - \tilde{A} \| \leqslant \delta$, $\| u - \tilde{u} \| \leqslant \delta$, *and let the inequality* $\| \tilde{A}v \| \geqslant \varkappa \| v \|$, *where* $\varkappa > 0$ *is a constant, be fulfilled for any* $v \in L(e_1, \ldots, e_{r-1})$, $r > 1$. *Then, for sufficiently small* $\delta > 0$,

$$\operatorname{diam} M_\delta \leqslant C_1 \delta^{1/2}, \quad C_1 > 0, \quad C_1 = \text{const},$$

where

$$M_\delta = \{x \mid x = p_0 + v, \; v \in L(e_1, \ldots, e_{r-1}), \quad \| \tilde{A}x - \tilde{u} \|^2 \leqslant \delta\}.$$

Proof. Indeed, for any $v \in L(e_1, \ldots, e_{r-1})$ such that $x \in M_\delta$ we have

$$\sqrt{\delta} \geqslant \| \tilde{A}(p_0 + v) - \tilde{u} \| = \| (\tilde{A} - A) p_0 + u - \tilde{u} + \tilde{A}v \|$$
$$\geqslant \| \tilde{A}v \| - \| (\tilde{A} - A) p_0 + u - \tilde{u} \| \geqslant \| \tilde{A}v - C_2\delta \geqslant \varkappa \| v \| - C_2\delta,$$
$$C_2 = \| p_0 \| + 1.$$

Consequently, $\| v \| \leqslant \frac{1}{\varkappa} (\delta^{1/2} + C_2\delta)$, i.e. $\operatorname{diam} M_\delta \leqslant C_1 \delta^{1/2}$.

Let $\{e_r, \ldots, e_n\}$ be a basis in the subspace $\operatorname{Ker} A = \{x \mid x \in E_n, \; Ax = 0\}$ and let $\{e_1, \ldots, e_{r-1}\}$ be a basis in the orthogonal complement $(\operatorname{Ker} A)^\perp$ of the kernel of the matrix A.

Theorem 2. *For sufficiently small* $\delta > 0$ *the estimate* $\| \tilde{x} - p_0 \| \leqslant C_3 \delta^{1/2}$ *takes place.*

Proof. For any $x \in (\operatorname{Ker} A)^\perp$ the inequality $\| Ax \| \geqslant \varkappa \| x \|$ holds, where $\varkappa$ is a positive constant not depending on x. Then for sufficiently small $\delta > 0$ the perturbed matrix $\tilde{A}$ satisfies the inequality

$$\| \tilde{A}x \| \geqslant \frac{\varkappa}{2} \| x \| \tag{5}$$

for any $x \in (\operatorname{Ker} A)^\perp$.

We note that for small disturbances δ the inequality (5) implies $\| A\mathbf{e}_i \| \geqslant \varkappa/2 > 0$ for all $\mathbf{e}_i$ $(i = 1, \ldots, r-1)$, i.e. the basic vectors of $(\operatorname{Ker} A)^{\perp}$ are retained in $(\operatorname{Ker} \tilde{A})^{\perp}$. For the vectors $\{\mathbf{e}_r, \ldots, \mathbf{e}_n\}$ there are two possible situations:

(a) $\tilde{A}\mathbf{e}_{r+i} = 0$ for all $i \in \{0, 1, \ldots, n-r\}$,

(b) $\tilde{A}\mathbf{e}_{r+i} \neq 0$ for some $i \in \{0, 1, \ldots, n-r\}$.

In the case (a) we have $\tilde{\mathbf{x}} \in M_\delta = \{\mathbf{x} \mid \mathbf{x} = \mathbf{p}_0 + \mathbf{v},\ \mathbf{v} \in L(\mathbf{e}_1, \ldots, \mathbf{e}_{r-1}),\ \| \tilde{A}\mathbf{x} - \tilde{\mathbf{u}} \|^2 \leqslant \delta\}$. By Lemma 2, $\operatorname{diam} M_\delta \leqslant C_1\delta^{1/2}$. Consequently, $\| \tilde{\mathbf{x}} - \mathbf{p}_0 \| \leqslant C_1\delta^{1/2}$.

In the case (b), let, for definiteness, $\tilde{A}\mathbf{e}_r \neq 0$ and $\tilde{A}\mathbf{e}_{r+i} = 0$ for $i = 1, \ldots, n-r$. Then $\tilde{\mathbf{x}} \in \tilde{M}_\delta = \{\mathbf{x} \mid \mathbf{x} = \mathbf{p}_0 + \omega,\ \omega \in L(\mathbf{e}_1, \ldots, \mathbf{e}_r),\ \| \tilde{A}\mathbf{x} - \tilde{\mathbf{u}} \|^2 \leqslant \delta\}$.

Let $\mathbf{q}$ be the orthogonal projection of the point $\tilde{\mathbf{x}}$ on the line $P_r = \{\mathbf{x} \mid \mathbf{x} = \mathbf{p}_0 + \gamma\mathbf{e}_r,\ \gamma \in E_1\}$. Then $\mathbf{q} = \mathbf{p}_0 + \gamma_q\mathbf{e}_r$. Let, for definiteness, $\gamma_q \geqslant 0$. Consider the point $\mathbf{z} = \mathbf{p}_0 - C_4\delta^{-1/2}\mathbf{e}_r$. Since $\| \tilde{A}\mathbf{e}_r \| = \| (A - \tilde{A})\,\mathbf{e}_r \| \leqslant \delta$, we have

$$\| \tilde{A}\mathbf{p}_0 - \tilde{\mathbf{u}} - A\mathbf{p}_0 + \mathbf{u} - C_4\delta^{-1/2}\tilde{A}\mathbf{e}_r \|$$

$$\leqslant \| A - \tilde{A} \| \, \| \mathbf{p}_0 \| + \| \mathbf{u} - \tilde{\mathbf{u}} \| + C_4\delta^{1/2} \leqslant \delta(\| \mathbf{p}_0 \| + 1) + C_4\delta^{1/2}.$$

Consequently, $\| \tilde{A}\mathbf{z} - \tilde{\mathbf{u}} \| \leqslant \delta^{1/2}$ for sufficiently small δ and $0 \leqslant C_4 < 1/2$, i.e. $\mathbf{z} \in \tilde{Y}$.

The vector $(\tilde{\mathbf{x}} - \mathbf{q})$ belongs to M_δ and is orthogonal to $\mathbf{e}_r$. By Lemma 2, $\operatorname{diam} M_\delta \leqslant C_1\delta^{1/2}$, and therefore

$$\| \tilde{\mathbf{x}} - \mathbf{q} \| \leqslant C_1\delta^{1/2}. \tag{6}$$

By virtue of the monotonicity of the minimizing sequence, the element $\mathbf{z}$ satisfies the inequality

$$\| \mathbf{x}_0 - \mathbf{z} \| > \| \tilde{\mathbf{x}} - \mathbf{z} \|. \tag{7}$$

Since the vector $(\mathbf{p}_0 - \mathbf{x}_0)$ is orthogonal to $\mathbf{e}_r$, we have

$$\| \mathbf{x}_0 - \mathbf{z} \|^2 = \| \mathbf{x}_0 - \mathbf{p}_0 \|^2 + 2\langle \mathbf{x}_0 - \mathbf{p}_0,\ \mathbf{p}_0 - \mathbf{z} \rangle + \| \mathbf{p}_0 - \mathbf{z} \|^2$$

$$= \| \mathbf{x}_0 - \mathbf{p}_0 \|^2 + \| \mathbf{p}_0 - \mathbf{z} \|^2,$$

i.e.

$$\| \mathbf{x}_0 - \mathbf{z} \| - \| \mathbf{p}_0 - \mathbf{z} \| = \frac{\| \mathbf{x}_0 - \mathbf{p}_0 \|^2}{\| \mathbf{x}_0 - \mathbf{z} \| + \| \mathbf{p}_0 - \mathbf{z} \|} \leqslant \frac{\| \mathbf{x}_0 - \mathbf{p}_0 \|^2}{\| \mathbf{p}_0 - \mathbf{z} \|} \leqslant C_5\delta^{1/2}. \tag{8}$$

The definition of the point $\mathbf{q}$ and the fact that $\mathbf{z} \in P_r$ imply that $\| \mathbf{q} - \mathbf{z} \| < \| \tilde{\mathbf{x}} - \mathbf{z} \|$. Thus, taking (7) and (8) into consideration,

we find

$$\| \mathbf{q} - \mathbf{z} \| < \| \mathbf{x}_0 - \mathbf{z} \| \leqslant \| \mathbf{p}_0 - \mathbf{z} \| + C_5 \delta^{1/2}. \tag{9}$$

Since the points $\mathbf{q}$, $\mathbf{p}_0$, and $\mathbf{z}$ lie in one line P_r, we see that, in view of the choice of $\mathbf{z}$, the inequality

$$\| \mathbf{q} - \mathbf{z} \| = \| \mathbf{q} - \mathbf{p}_0 \| + \| \mathbf{z} - \mathbf{p}_0 \|$$

takes place, and therefore

$$\| \mathbf{q} - \mathbf{p}_0 \| + \| \mathbf{z} - \mathbf{p}_0 \| < \| \mathbf{p}_0 - \mathbf{z} \| + C_5 \delta^{1/2},$$

i.e. $\| \mathbf{q} - \mathbf{p}_0 \| < C_5 \delta^{1/2}$. With regard to (5), we obtain

$$\| \tilde{\mathbf{x}} - \mathbf{p}_0 \| \leqslant \| \tilde{\mathbf{x}} - \mathbf{q} \| + \| \mathbf{q} - \mathbf{p}_0 \| < C_3 \delta^{1/2},$$
$$C_3 = C_1 + C_5. \quad \triangle$$

In conclusion we note that the sequence $\{\mathbf{x}_k\}$ minimizing the function $\tilde{\varphi}(\mathbf{x}) = \| \tilde{A}\mathbf{x} - \tilde{\mathbf{u}} \|^2$ according to the suggested scheme belongs to the set $M = \mathbf{x}_0 + (\operatorname{Ker} \tilde{A}^T \tilde{A})^{\perp}$, which results from the translation of the subspace $(\operatorname{Ker} \tilde{A}^T \tilde{A})^{\perp}$ by the vector $\mathbf{x}_0$, where the quadratic function $\tilde{\varphi}(\mathbf{x}) = \| \tilde{A}\mathbf{x} - \tilde{\mathbf{u}} \|^2$ is strongly convex. Therefore for the function $\tilde{\varphi}(\mathbf{x})$ the estimate for the rate of convergence (see Sec. 9.5.3) $\| \mathbf{x}_k - \tilde{\mathbf{x}} \|^2 \leqslant \lambda^{-1} \exp(-Ck)$, $C = \text{const} > 0$, holds, where λ is the minimum nonzero eigenvalue of the matrix $\tilde{A}^T \tilde{A}$ and $\tilde{\mathbf{x}} = \arg\min \tilde{\varphi}(\mathbf{x})$. Consequently,

$$\| \mathbf{x}_k - \mathbf{p}_0 \| \leqslant \| \mathbf{x}_k - \tilde{\mathbf{x}} \| + \| \tilde{\mathbf{x}} - \mathbf{p}_0 \| \lambda^{-1/2} \exp(-Ck/2) + C_3 \delta^{1/2}.$$

Thus, the gradient method with a special choice of the step-length possesses the property of self-regularization. It is obvious that the number of steps k required to attain an accuracy of $O(\delta^{1/2})$ must be of the order of $|\ln \delta|$.

BIBLIOGRAPHY

1. Bazaraa, M.S. and Jarvis, J.J., *Linear Programming and Network Flows*, Wiley and Sons, New York, 1977
2. Bazaraa, M.S. and Shetty C.M., *Nonlinear Programming. Theory and Algorithms*, Wiley and Sons, New York, 1979
3. Bertsekas, D.P., *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982
4. Dantzig, G.B., *Linear Programming and Extensions*, Princeton University Press, New Jersey, 1963
5. Fadeev, D.K. and Fadeeva, V.N., *Computational Methods of Linear Algebra*, Freeman, San Francisco, California, 1963
6. Fiacco, A.V. and McCormick, G.P., *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*, Wiley, New York, 1968
7. Ortega, J.M. and Rheinboldt, W.C., *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970
8. Polak, E., *Computational Methods in Optimization: a Unified Approach*, Academic Press, New York, 1971
9. Rockafellar, R.T., *Convex Analysis*, Princeton Univ. Press, Princeton, New Jersey, 1970
10. Zangwill, W.I., *Nonlinear Programming*, Prentice-Hall, Englewood, Cliffs., New Jersey, 1969

NAME INDEX

Arrow, K.J. 276

Banach, S. 286, 287, 290
Bazaraa, M.S. 301
Bereznev, V.A. 9
Bertsekas, D.P. 301
Binet, J.P.M. 156, 158
Bunyakovskiĭ, V.Ya. 24, 203

Cauchy, A.L. 24, 203

Dantzig, G.B. 201
Davidon, W.C. 200
Denisov, D.V. 9

Fadeev, D.K. 301
Fadeeva, V.N. 301
Farkas, J. 34
Fiacco, A.V. 301
Fibonacci, L. 154, 158
Fletcher, R. 200

Hausdorff, F. 286
Hurwicz, L. 276

Javis, J.J. 301

Kuhn, H.W. 58, 65, 67

Lagrange, J.L. 15, 62, 247, 248, 269
Lipschitz, R.O.S. 168, 228
Lyusternik, L.A. 286

McCormick, G.P. 301

Ortega, J.M. 301

Polak, E. 301
Powell, M.J.D. 200

Rheinboldt, W.C. 301
Rockafellar, R.F. 301

Shetty, C.M. 301

Taylor, B. 206, 289, 292
Tikhonov, A.N. 142
Tret'yakov, A.A. 9
Tucker, A.W. 58, 65, 67

Zangwill, W.I. 301

SUBJECT INDEX

Accelerated Newton's method 205
Active constraints 50
Allocation problem 20
Almost certain convergence 249, 262
Arrow-Hurwicz method 276
Artificial basis 86, 94f
Assignment problem 18

Basic components 89
Basic mathematical programming problem 48
Binet's formula 156, 158
Block problem 117
Boundary point 25
Boundary of a set 25

Cauchy-Bunyakovskii inequality 24, 203
Class $C^{1,1}(X)$ 193
Closed interval 22
Closed transportation problem 18
Column vector 21
Concave function 37
Condition matrix 69
Condition vector 69
Conditional probability 249
Cone 33
Conjugate directions method 187
Conjugate directions relative to a matrix 214
Constraints 11, 49
Constraint vector 69
Controllable phenomena 10
Control parameters 11
Convergence in probability 249, 261, 263
Convergent sequence 25
Convex combination 25
Convex function 37
Convex programming problems 14, 19, 49
Convex set 25
Coordinates of a point 24
Coordinate-wise descent method 187, 244ff
Correct (correctly posed) problem 138
Cubic approximation method 164f
Cyclic coordinate-wise descent method 201
Cycling 98

Davidon-Fletcher-Powell method 200, 201f
Decision making 10
Decomposition method 116ff
Degeneracy in extremal problems 284
Descent methods 186ff
Diet problem 16
Discrepancy vector 113
Distance between a point and a set 25
Distance to the boundary 235
Dual basis of the primal problem 104
Duality 66ff
Dual problem 67f, 83
Dual simplex method 104ff
Dual variables 67
Duplication method 188

Elementary matrix 103
Empty set 21
ε-neighborhood 25
ε-problem 99
Equivalent extremal problems 68
Euclidean norm 21, 24
Euclidean space 21, 24
Extremal problems 11, 117ff
Extreme point 31

Farkas theorem 34
Feasible directions 49, 234
Feasible point 49
Feasible region 49
Fibonacci method 154, 158
Fibonacci numbers 156
First duality theorem 71
First-order methods 198

γ-majorizable set 135
Generalized gradient descent method 199f
Generalized Lyusternik theorem 286, 292
Generalized principle of contracting mapping 285, 287
Golden section method 154, 158
Gradient descent methods 187, 198f, 269, 300
Gradient projection method 224ff, 269

Half-open interval 22
Hausdorff's distance between sets 286
Hybrid methods 256

Incorrect (incorrectly posed) problems 139
Integer programming problems 15
Interior point of a set 25
Interior of a set 21
Intersection of sets 21
Iterative penalty function method 296

Jensen's inequality 39

Kuhn-Tucker point 67
Kuhn-Tucker theorem 58, 65, 273ff, 277

Lagrange function 62, 247f
Lagrange multiplier method 15, 62, 269
Linearly non-constant functions 179
Linear programming problems 14, 85
Lipschitz condition 168
Lipschitz constant 228

Mathematical model 11
Mathematical programming 10, 14
Mean value theorem 286
Method
of conditional gradient 222, 229ff
of conjugate directions 209ff
of feasible directions 234
of finding a feasible point 244ff
of finding an initial extreme point 94ff
of projecting a random direction 267
of simultaneous solution of the primal and dual problems 111ff
of tangents 165ff
M-method 96f
Modified golden section method 163
Modified Lagrange function method 270ff, 276ff
Modified Newton's method 205
Modified simplex method 102
Multiextremum problems 15, 21
Multiplicative representation of a matrix 103

Newton's method 202ff
Nonbasic components 89
Non-degenerate extreme point 86
Nonlinear programming problems 14
Normal solution 142

Objective function 11
One-dimensional minimization problems 151ff
Open mapping theorem of Banach 286f, 290
Open interval 22
Open transportation problem 18
Operations research 10
Optimal point 41, 49

Parabolic method 163f
Parameter of strong convexity 45, 255
Penalty 123
Penalty function method 122ff, 126, 131ff, 293ff
Perturbation method 98f
Point in a Euclidean space 24
p-regularity of constraints 285
Primal problem 66
Problems with two-sided constraints 101
Product of matrices 22
Projection 26
*p*th-order necessary and sufficient conditions for optimality 291ff

Quadratic programming problems 14
Qualitative model 11
Quasi-Newton type methods 209ff

Random coordinate-wise descent 255
Random descent method 249ff, 256ff
Rate of convergence 195, 207, 227, 232, 254, 281
Recurrence relations of the simplex method 93f
Reduction of the basic linear programming problem to the canonical form 86
Regularity condition 54
Regularization method 140ff, 145ff
Regularization parameters 156
Reinitialization 202
Relaxation process 173
Row vector 21

Saddle point 60
Scalar product 21
Second duality theorem 72
Second-order methods 198
Selection of a column to be introduced in the basis 89f
Selection of a column to be removed from the basis 90ff
Self-regularization 300
Separating hyperplane 30
Separation theorem 28
Set of stationary points 193
Set-theoretical difference 21
Simplex method 89ff, 101
Slater's regularity condition 54, 237
Solution to a mathematical programming problem 49
Stability of minimization methods 172f
Stability properties of the gradient method 296ff
Stationary point 49
Steepest descent method 200
Steplength 188
Stochastic programming 14
Strictly convex function 37
Strictly quasi-convex function 179
Strictly unimodal function 152
Strongly convex function 44
Strongly lowering sequence 179
Sufficient conditions for optimality 58ff
Supporting hyperplane 28

Tangent cone to a set 285
Tangent vector to a set 285
Taylor's formula 206, 289, 292
Theorem of addition of probability 253
Theorem on estimates 182
Theorem of total probability 252
Tikhonov's regularization method 142ff
Transportation problem 14, 16
Transpose of a matrix 22

Uniform probability distribution 256
Union of sets 21

Variable scale method 200

Weakly correct problems 138

Zero-order methods 214ff, 198